This CVPR 2020 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Ricard Durall[1,3], Margret Keuper[2], Janis Keuper[1,4]

[1]Competence Center High Performance Computing, Fraunhofer ITWM, Kaiserslautern, Germany
[2]Data and Web Science Group, University of Mannheim, Germany
[3]IWR, University of Heidelberg, Germany
[4]Institute for Machine Learning and Analytics, Offenburg University, Germany

Abstract

Generative convolutional deep neural networks, e.g. *popular GAN architectures, are relying on convolution based up-sampling methods to produce non-scalar outputs like images or video sequences. In this paper, we show that common up-sampling methods,* i.e. *known as up-convolution or transposed convolution, are causing the inability of such models to reproduce spectral distributions of natural training data correctly. This effect is independent of the underlying architecture and we show that it can be used to easily detect generated data like deepfakes with up to 100% accuracy on public benchmarks. To overcome this drawback of current generative models, we propose to add a novel spectral regularization term to the training optimization objective. We show that this approach not only allows to train spectral consistent GANs that are avoiding high frequency errors. Also, we show that a correct approximation of the frequency spectrum has positive effects on the training stability and output quality of generative networks.*

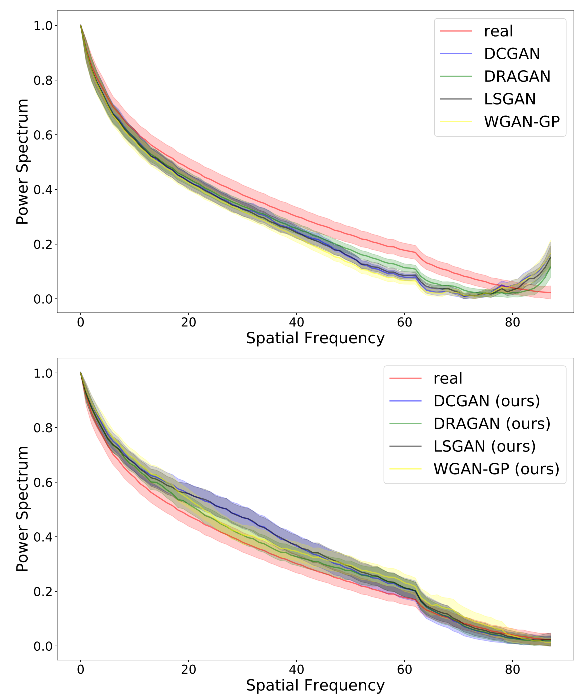


Figure 1: Common up-convolution methods are inducing heavy spectral distortions into generated images. The **top** figure shows the statistics (mean and variance) after azimuthal integration over the power-spectrum (see Section 2.1) of real and GAN generated images. Evaluation on the *CelebA* [34] data set, here all GANs (DCGAN [47], DRAGAN [32], LSGAN [37], WGAN-GP [20]) are using "transposed convolutions" (see Section 2.2) for up-sampling.
Bottom: Results of the same experiments as above, adding our proposed spectral loss during GAN training.

1. Introduction

Generative convolutional deep neural networks have recently been used in a wide range of computer vision tasks: generation of photo-realistic images [29, 6], image-to-image [45, 26, 61, 9, 42, 30] and text-to-image translations [48, 11, 58, 59], style transfer [27, 60, 61, 25], image inpainting [45, 54, 33, 26, 56], transfer learning [5, 10, 15] or even for training semantic segmentation tasks [35, 53], just to name a few.

The most prominent generative neural network architectures are *Generative Adversarial Networks (GAN)* [18] and *Variational Auto Encoders (VAE)* [46]. Both basic approaches try to approximate a latent-space model of the underlying (image) distributions from training data samples. Given such a latent-space model, one can draw new (artificial) samples and manipulate their semantic properties in various dimensions. While both GAN and VAE approaches have been published in many different variations, e.g. with different loss functions [18, 4, 20], different latent space constraints [41, 13, 13, 21, 30] or various deep neural network (DNN) topologies for the generator networks [47, 43], all of these methods have to follow a basic data generation principle: they have to transform samples from a low dimensional (often 1D) and low resolution latent space to the high resolution (2D image) output space. Hence, these generative neural networks must provide some sort of (learnable) up-scaling properties.

While all of these generative methods are steering the learning of their model parameters by optimization of some loss function, most commonly used losses are focusing exclusively on properties of the output image space, e.g. using convolutional neural networks (CNN) as discriminator networks for the implicit loss in an image generating GAN. This approach has been shown to be sufficient in order to generate visually sound outputs and is able to capture the data (image) distribution in image-space to some extent. However, it is well known that up-scaling operations notoriously alter the spectral properties of a signal [28], causing high frequency distortions in the output.

In this paper, we investigate the impact of up-sampling techniques commonly used in generator networks. The top plot of Figure 1 illustrates the results of our initial experiment, backing our working hypotheses that current generative networks fail to reproduce spectral distributions. Figure 1 also shows that this effect is independent of the actual generator network.

1.1. Related Work

1.1.1 Deepfake Detection

We show the practical impact of our findings for the task of *Deepfake detection*. The term *deepfake* [22, 8] describes the recent phenomenon of people misusing advances in artificial face generation via deep generative neural networks [7] to produce fake image content of celebrities and politicians. Due to the potential social impact of such fakes, *deepfake detection* has become a vital research topic of its own. Most approaches reported in the literature, like [38, 3, 57], are themselves relying on CNNs and thus require large amounts of annotated training data. Likewise, [24] introduces a deep forgery discriminator with a contrastive loss function and [19] incorporates temporal domain information by employing Recurrent Neural Networks (RNNs) on top of CNNs.

1.1.2 GAN Stabilization

Regularizing GANs in order to facilitate a more stable training and to avoid mode collapse has recently drawn some attention. While [40] stabilize GAN training by unrolling the optimization of the discriminator, [50] propose regularizations via noise as well as an efficient gradient-based approach. A stabilized GAN training based on octave convolutions has recently been proposed in [16]. None of these approaches consider the frequency spectrum for regularization. Yet, very recently, band limited CNNs have been proposed in [17] for image classification with compressed models. In [55], first observations have been made that hint towards the importance of the power spectra on model robustness, again for image classification. In contrast, we propose to leverage observations on the GAN generated frequency spectra for training stabilization.

1.2. Contributions

The contributions of our work can be summarized as follows:

- We experimentally show the inability of current generative neural network architectures to correctly approximate the spectral distributions of training data.
- We exploit these spectral distortions to propose a very simple but highly accurate detector for generated images and videos, i.e. a *DeepFake* detector that reaches up to 100% accuracy on public benchmarks.
- Our theoretical analysis and further experiments reveal that commonly used up-sampling units, i.e. up-convolutions, are causing the observed effects.
- We propose a novel **spectral regularization** term which is able to compensate spectral distortions.
- We also show experimentally that using spectral regularization in GAN training leads to more stable models and increases the visual output quality.

The remainder of the paper is organized in as follows: Section 2 introduces common up-scaling methods and analyzes their negative effects on the spectral properties of images. In Section 3, we introduce a novel spectral-loss that allows to train generative networks that are able to compensate the up-scaling errors and generate correct spectral distributions. We evaluate our methods in Section 4 using current architectures on public benchmarks.

2. The Spectral Effects of Up-Convolutions

2.1. Analyzing Spectral Distributions of Images using Azimuthal Integration over the DFT Power Spectrum

In order to analyze effects on spectral distributions, we rely on a simple but characteristic 1D representation of the

Fourier power spectrum. We compute this spectral representation from the discrete Fourier Transform $\mathcal{F}$ of 2D (image) data I of size $M \times N$,

$$\mathcal{F}(I)(k,\ell) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} e^{-2\pi i \cdot \frac{jk}{M}} e^{-2\pi i \cdot \frac{j\ell}{N}} \cdot I(m,n), \quad (1)$$
$$\text{for} \quad k = 0,\ldots,M-1, \quad \ell = 0,\ldots,N-1,$$

via azimuthal integration over radial frequencies ϕ

$$AI(\omega_k) = \int_0^{2\pi} \|\mathcal{F}(I)\left(\omega_k \cdot \cos(\phi), \omega_k \cdot \sin(\phi)\right)\|^2 \mathrm{d}\phi$$
$$\text{for} \quad k = 0,\ldots,M/2-1\,, \quad (2)$$

assuming square images[1]. Figure 2 gives a schematic impression of this processing step.

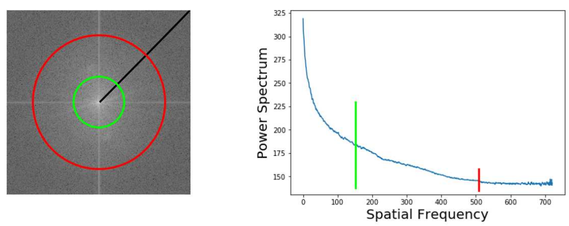


Figure 2: Example for the azimuthal integral (AI). (Left) 2D Power Spectrum of an image. (Right) 1D Power Spectrum: each frequency component is the radial integral over the 2D spectrum (red and green examples).

2.2. Up-convolutions in generative DNNs

Generative neural architectures like GANs produce high dimensional outputs, *e.g.* images, from very low dimensional latent spaces. Hence, all of these approaches need to use some kind of up-scaling mechanism while propagating data through the network. The two most commonly used up-scaling techniques in literature and popular implementations frameworks (like TensorFlow [2] and PyTorch [44]) are illustrated in Figure 3: **up-convolution by interpolation** (up+conv) and **transposed convolution** (transconv) . We use a very simple auto encoder (AE) setup (see Figure 4) for an initial investigation of the effects of up-convolution units on the spectral properties of 2d images after up-sampling. Figure 5 shows the different, but massive impact of both approaches on the frequency spectrum. Figure 6 gives a qualitative result for a reconstructed image and shows that the mistakes in the frequency spectrum are relevant for the visual appearance.

[1]$\rightarrow M = N$. We are aware that this notation is abusive, since $\mathcal{F}(I)$ is discrete. However, fully correct discrete notation would only over complicated a side aspect of our work. A discrete implementations of AI is provided on `https://github.com/cc-hpc-itwm/UpConv`.

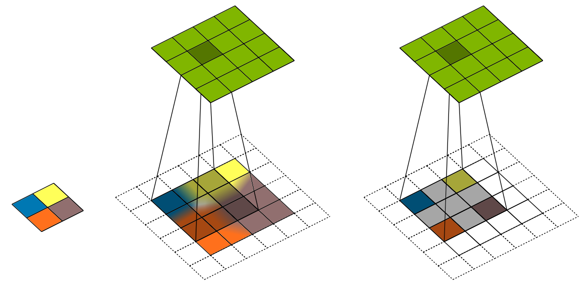


Figure 3: Schematic overview of the two most common up-convolution units. **Left: low resolution input image** (here 2×2); **Center: up-convolution by interpolation** (up+conv) - the input is scaled via interpolation (bi-linear or nearest neighbor) and then convolved with a standard learnable filter kernel (of size 3×3) to form the 5x5 output (green), **Right: transposed convolution** (transconv) - the input is padded with a "bed of nails" scheme (gray grid points are zero) and then convolved with a standard filter kernel to form the 5×5 output (green).

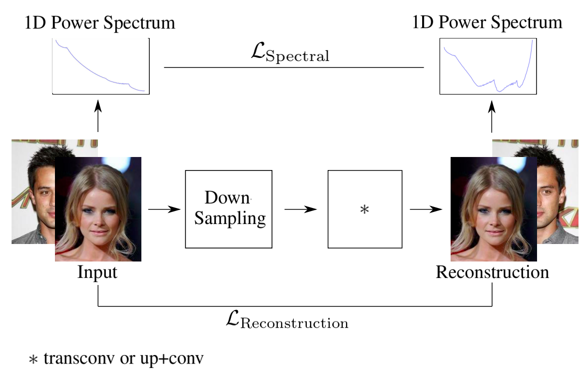


Figure 4: Schematic overview of the simple auto encoder (AE) setup used to demonstrate the effects of up-convolutions in Figure 5, using only a standard MSE reconstruction loss (bottom) to train the AE on real images. We down-scale the input by a factor of two and then use the different up-convolution methods to reconstruct the original image size. In Section 3 we use the additional spectral loss (top) to compensate the spectral distortions (see Figure 7)

.

2.3. Theoretical Analysis

For the theoretic analysis, we consider, without loss of generality, the case of a one-dimensional signal a and its discrete Fourier Transform $\hat{a}$

$$\hat{a}_k = \sum_{j=0}^{N-1} e^{-2\pi i \cdot \frac{jk}{N}} \cdot a_j, \quad \text{for} \quad k = 0,\ldots,N-1. \quad (3)$$

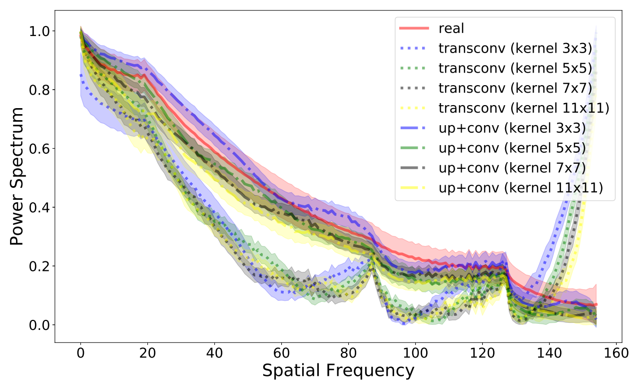


Figure 5: Effects of single up-convolution units (setup see Figure 4) on the frequency spectrum (azimuthal integral) of the output images. Both up-convolution methods have massive effects on the spectral distributions of the outputs. Transposed convolutions add large amounts high frequency noise while interpolation based methods (up+conv) are lacking high frequencies.

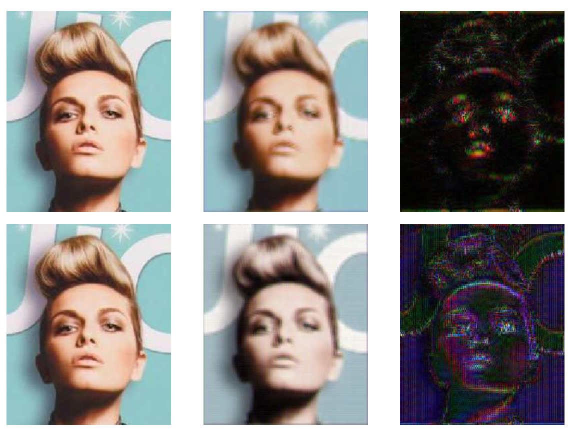


Figure 6: Effects of spectral distortions on the image outputs in our simple AE setting. **Left:** original image; **Center:** AE output image; **Right:** filtered difference image . The **top row** shows the blurring effect of missing high frequencies in the (up+conv) case; **Bottom row** shows the high frequency artifacts induces by (transconv).

If we want to increase a's spatial resolution by factor 2, we get

$$\hat{a}^{up}_{\bar{k}} = \sum_{j=0}^{2\cdot N-1} e^{-2\pi i \cdot \frac{j\bar{k}}{2\cdot N}} \cdot a^{up}_j \tag{4}$$

$$= \sum_{j=0}^{N-1} e^{-2\pi i \cdot \frac{2\cdot j\bar{k}}{2\cdot N}} \cdot a_j + \sum_{j=0}^{N-1} e^{-2\pi i \cdot \frac{2\cdot(j+1)\bar{k}}{2\cdot N}} \cdot b_j, \tag{5}$$

for $\quad \bar{k} = 0, \ldots, 2N-1.$

where $b_j = 0$ for "bed of nails" interpolation (as used by *transconv*) and $b_j = \frac{a_{j-1}+a_j}{2}$ for bi-linear interpolation (as used by *up+conv*).

Let us first consider the case of $b_j = 0$, *i.e.* "bed of nails" interpolation. There, the second term in Eq. (6) is zero. The first term is similar to the original Fourier Transform, yet with the parameter k being replaced by $\bar{k}$. Thus, increasing the spatial resolution by a factors of 2 leads to a scaling of the frequency axes by a factor of $\frac{1}{2}$. Let us now consider the effect from a sampling theory based viewpoint. It is

$$\hat{a}^{up}_{\bar{k}} = \sum_{j=0}^{2\cdot N-1} e^{-2\pi i \cdot \frac{j\bar{k}}{2\cdot N}} \cdot a^{up}_j \tag{6}$$

$$= \sum_{j=0}^{2\cdot N-1} e^{-2\pi i \cdot \frac{j\bar{k}}{2\cdot N}} \cdot \sum_{t=-\infty}^{\infty} a^{up}_j \cdot \delta(j-2t) \tag{7}$$

since the point-wise multiplication with the Dirac impulse comb only removes values for which $a^{up} = 0$. Assuming a periodic signal and applying the convolution theorem [31], we get

$$(7) = \frac{1}{2} \cdot \sum_{t=-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} e^{-2\pi i \cdot \frac{j\bar{k}}{2\cdot N}} a^{up}_j \right) \left(\bar{k} - \frac{t}{2} \right), \tag{8}$$

which equals to

$$\frac{1}{2} \cdot \sum_{t=-\infty}^{\infty} \left(\sum_{j=-\infty}^{\infty} e^{-2\pi i \cdot \frac{j\bar{k}}{N}} \cdot a_j \right) \left(\bar{k} - \frac{t}{2} \right) \tag{9}$$

by Eq. (6). Thus, the "bed of nails upsampling" will create high frequency replica of the signal in $\hat{a}^{up}$. To remove these frequency replica, the upsampled signal needs to be smoothed appropriately. All observed spatial frequencies beyond $\frac{N}{2}$ are potential upsampling artifacts. While it is obvious from a theoretical point of view, we also demonstrate practically in Figure 8 that the correction of such a large frequency band is (assuming medium to high resolution images) is not possible with the commonly used 3×3 convolutional filters.

In the case of bilinear interpolation, we have $b_j = \frac{a_{j-1}+a_j}{2}$ in Eq. (6), which corresponds to an average filtering of the values of a adjacent to b_j. This is equivalent to a point-wise multiplication of a^{up} spectrum $\hat{a}^{up}$ with a sinc function by their duality and the convolution theorem, which suppresses artificial high frequencies. Yet, the resulting spectrum is expected to be overly low in the high frequency domain.

3. Learning to Generate Correct Spectral Distributions

The experimental evaluations of our findings in the previous section and their application to detect generated content (see Section 4.1), raise the question if it would be possible to correct the spectral distortion induced by the up-convolution units used in generative networks. After all, usual network topologies contain learnable convolutional filters which follow the up-convolutions and potentially could correct such errors.

3.1. Spectral Regularization

Since common generative network architectures are mostly exclusively using image-space based loss functions, it is not possible to capture and correct spectral distortions directly. Hence, we propose to add an additional spectral term to the generator loss:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\text{Generator}} + \lambda \cdot \mathcal{L}_{\text{Spectral}} \,, \tag{10}$$

where λ is the hyper-parameter that weights the influence of the spectral loss. Since we are already measuring spectral distortions using azimuthal integration AI (see Eq. (2)), and AI is differentiable, a simple choice for $\mathcal{L}_{\text{Spectral}}$ is the **binary cross entropy** between the generated output AI^{out} and the mean AI^{real} obtained from real samples:

$$\mathcal{L}_{\text{Spectral}} := -\frac{1}{(M/2-1)} \sum_{i=0}^{M/2-1} AI_i^{real} \cdot \log(AI_i^{out}) + (1 - AI_i^{real}) \cdot \log(1 - AI_i^{out}) \tag{11}$$

Notice that M is the image size and we use normalization by the 0^{th} coefficient (AI_0) in order to scale the values of the azimuthal integral to $[0, 1]$.

The effects of adding our spectral loss to the AE setup from Section 2.2 for different values of λ are shown in Figure 7. As expected based on our theoretical analysis in sec. 2.3, the observed effects can not be corrected by a single, learned 3×3 filter, even for large values λ. We thus need to reconsider the architecture parameters.

3.2. Filter Sizes on Up-Convolutions

In Figure 8, we evaluate our spectral loss on the AE from Section 2.2 with respect to filter sizes and the number of convolutional layers. We consider varying decoder filter sizes from 3×3 to 11×11 and 1 or 3 convolutional layers. While the spectral distortions from the up-sampling can not be removed with a single and even not with three 3×3 convolutions, it can be corrected by the proposed loss when more, larger filters are learned.

4. Experimental Evaluation

We evaluate the findings of the previous sections in three different experiments, using prominent GAN architectures on public face generation datasets. Section 4.1 shows that common face generation networks produce outputs with strong spectral distortions which can be used to detect artificial or "fake" images. In Section 4.2, we show that our spectral loss is sufficient to compensate artifacts in the frequency domain of the same data. Finally, we empirically show in Section 4.3 that spectral regularization also has positive effects on the training stability of GANs.

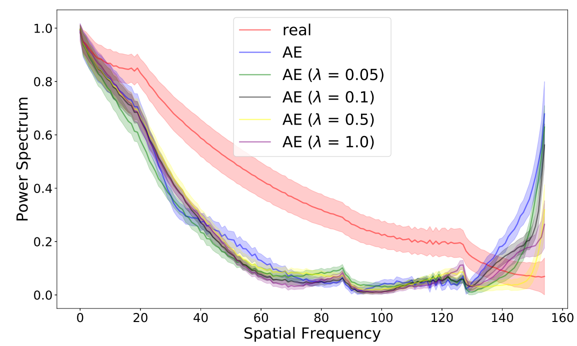


Figure 7: Auto encoder (AE) results with spectral loss by λ. Even if the spectral loss has a high weight, spectral distortions can not be corrected with a single 3×3 convolutional layer. This result is in line with the findings from Section 2.3.

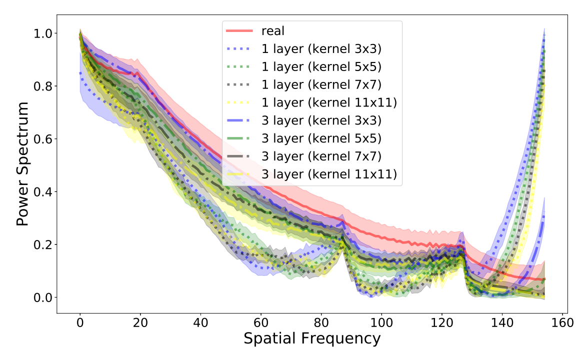


Figure 8: AE results with spectral loss by filter size of the convolution following the up-sampling step. The result heavily depends on the chosen filter size and number of convolutional layers. With three 5×5 convolutional filters available, the AE can greatly reduce spectral distortions using the proposed spectral loss.

4.1. Deepfake Detection

In this section, we show that the spectral distortions caused by the up-convolutions in state of the art GANs can be used to easily identify "fake" image data. Using only a small amount of annotated training data, or even an unsupervised setting, we are able to detect generated faces from public benchmarks with almost perfect accuracy.

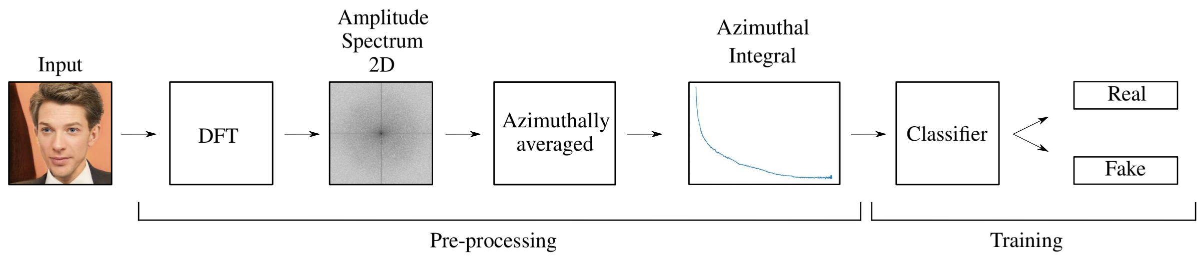


Figure 9: Overview of the processing pipeline of our approach. It contains two main blocks, a feature extraction block using DFT and a training block, where a classifier uses the new transformed features to determine whether the face is real or not. Notice that input images are converted to grey-scale before DFT.

4.1.1 Benchmarks

We evaluate our approach on three different data sets of facial images, providing annotated data with different spacial resolutions:

- *FaceForensics++* [49] contains a DeepFake detection data set with 363 original video sequences of 28 paid actors in 16 different scenes, as well as over 3000 videos with face manipulations and their corresponding binary masks. All videos contain a trackable, mostly frontal face without occlusions which enables automated tampering methods to generate realistic forgeries. The resolution of the extracted face images varies, but is usually around $80 \times 80 \times 3$ pixels.
- The CelebFaces Attributes (*CelebA*) dataset [34] consists of 202,599 celebrity face images with 40 variations in facial attributes. The dimensions of the face images are $178 \times 218 \times 3$, which can be considered to be a medium resolution in our context.
- In order to evaluate high resolution $1024 \times 1024 \times 3$ images, we provide the new *Faces-HQ* [2] data set, which is a annotated collection of 40k publicly available images from *CelebA-HQ* [29], *Flickr-Faces-HQ* dataset [30], 100K Faces project [1] and *www.thispersondoesnotexist.com*.

4.1.2 Method

Figure 9 illustrates our simple processing pipeline, extracting spectral features from samples via azimuthal integration (see Figure 2) and then using a basic SVM [51] classifier[3] for supervised and K-Means [36] for unsupervised fake detection. For each experiment, we randomly select training sets of different sizes and use the remaining data for testing. Training and test sets are equally balanced in their according class labels. All reported results are mean values of ten independent experiments.

In order to handle input images of different sizes, we normalize the 1D power spectrum by the 0^{th} coefficient and scale the resulting 1D feature vector to a fixed size.

4.1.3 Results

Figure 10 shows that real and "fake" faces form well delineated clusters in the high frequency range of our spectral feature space. The results of the experiments in Table 1 confirm that the distortions of the power spectrum, caused by the up-sampling units, are a common problem and allow an easy detection of generated content. This simple indicator even outperforms complex DNN based detection methods using large annotated training sets[4].

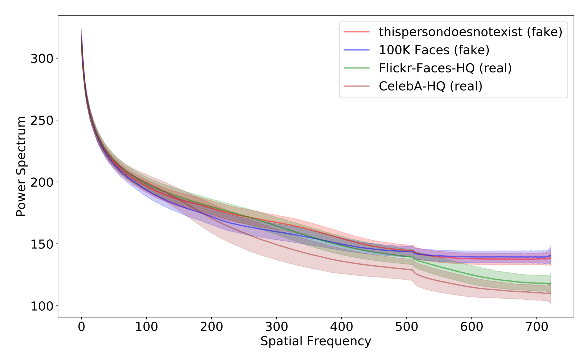


Figure 10: AI (1D power spectrum) statistics (mean and variance) of 1000 samples from each Faces-HQ sub-dataset. Clearly, real and "fake" images can be distinguished by their AI representation.

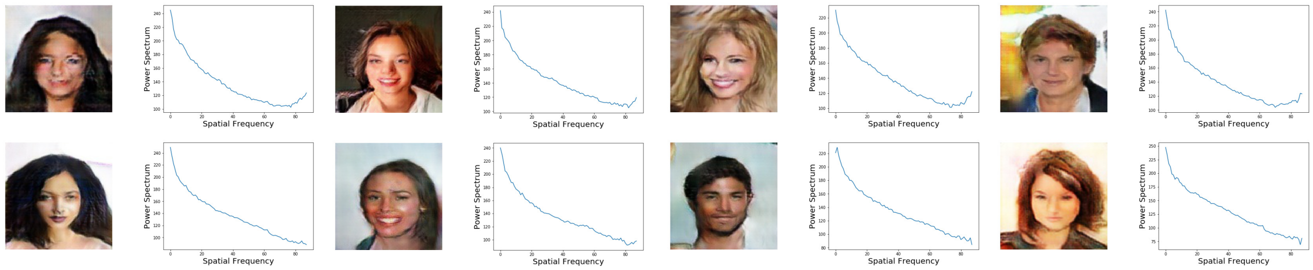
4.2. Applying Spectral Regularization

In this section, we evaluate the effectiveness of our regularization approach on the *CelebA* benchmark, as in the

[2]Faces-HQ data has a size of 19GB. Download: https://cutt.ly/6enDLYG. Also refer to [14].

[3]SVM hyper-parameters can be found in the source code

[4]Note: results of all other methods as reported by [57]. The direct comparison of methods might be biased since [57] used the same real data but generated the fake data independently with different GANs.



(a) DCGAN. (b) DRAGAN. (c) LSGAN. (d) WGAN.

Figure 11: Samples from the different types of GAN and their 1D Power Spectrum. **Top row:** samples produced by standard topologies. **Bottom row:** samples produced by standard topologies together with our spectral regularization technique.

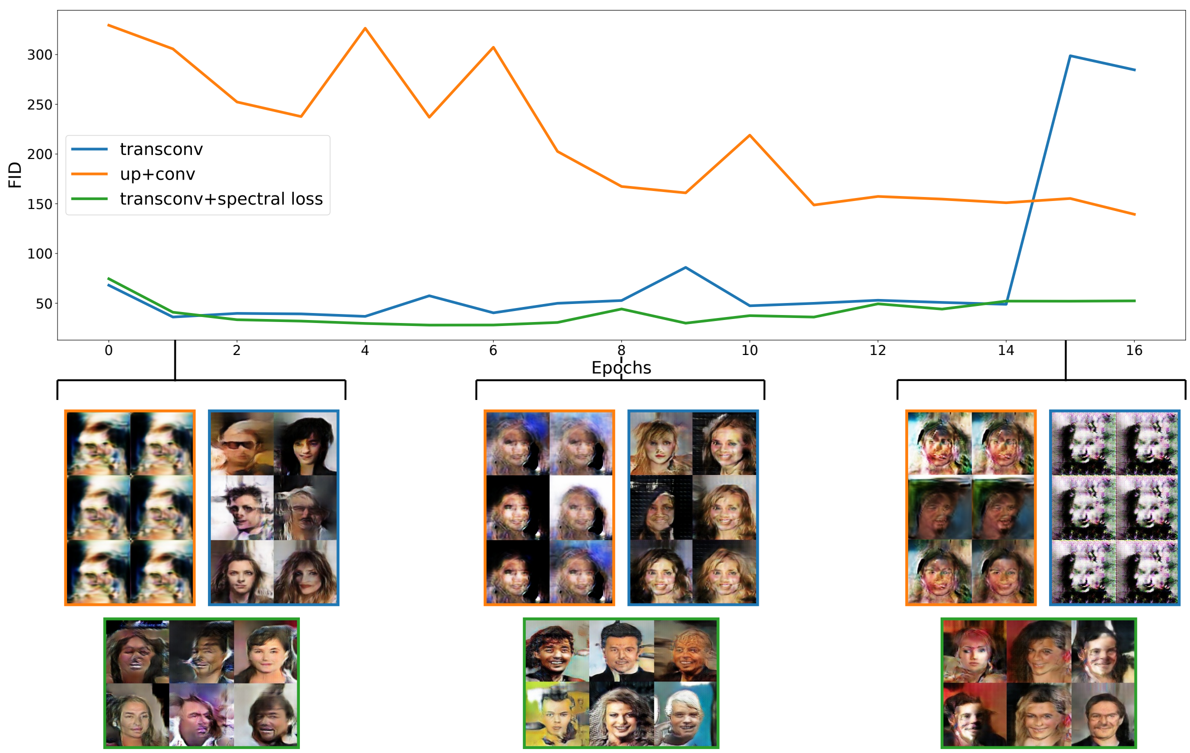


Figure 12: Correlation between FID values and GAN outputs for a DCGAN baseline on CelebA through out a training run. Low FID scores correspond to diverse but visually sound face image outputs. High FID scores indicate poor quality outputs and "mode collapse" scenarios where all generated images are bound to a very narrow sub-space of the original distribution.

experiment before. Based our theoretic analysis (see Section 2.3) and first AE experiments in Section 3, we extend existing GAN architectures in two ways: first, we add a spectral loss term (see Eq. (11)) to the generator loss. We use 1000 unannotated real samples from the data set to estimate AI^{real}, which is needed for the computation of the spectral loss (see Eq. (11)). Second, we change the convolution layers after the last up-convolution unit to three filter layers with kernel size 5×5. The bottom plot of Figure1 shows the results for this experiment in direct comparison to the original GAN architectures. Several qualitative results produced without and with our proposed regularization are given in Figure 11.

4.3. Positive Effects of Spectral Regularization

By regularizing the spectrum, we achieve the direct benefit of producing synthetic images that not only look realistic, but also mimic the behaviour in the frequency domain. In this way, we are one step closer to sample images from the real distribution. Additionally, there is an interesting side-effect of this regularization. During our experiments, we noticed that GANs with a spectral loss term appear to

data set	80% (train) - 20% (test)			
	method	# samples	supervised	unsupervised
Faces-HQ	ours	1000	100%	**82%**
Faces-HQ	ours	100	100%	81%
Faces-HQ	ours	20	**100%**	75%
CelebA	ours	2000	**100%**	**96%**
CelebA	[57]	100000	99.43%	-
CelebA	[39]	100000	86.61%	-
FaceForensics++	ours^A	2000	87%	-
FaceForensics++	ours^B	2000	**90%**	-

Table 1: Test accuracy. Our methods use SVM (supervised) and k-means (unsupervised) under different data settings. A) Evaluated on single frames. B) Accuracy on full video sequences via majority vote of single frame detections.

be much more stable in terms of avoiding "mode-collapse" [18] and better convergence. It is well known that GANs can suffer from challenging and unstable training procedures and there is little to no theory explaining this phenomenon. This makes it extremely hard to experiment with new generator variants, or to employ them in new domains, which drastically limits their applicability.

In order to investigate the impact of spectral regularization on the GAN training, we conduct a series of experiments. By employing a set of different baseline architectures, we assess the stability of our spectral regularization, providing quantitative results on the CelebA dataset. Our evaluation metric is the *Fréchet Inception Distance* (FID) [23], which uses the Inception-v3 [52] network pre-trained on ImageNet [12] to extract features from an intermediate layer.

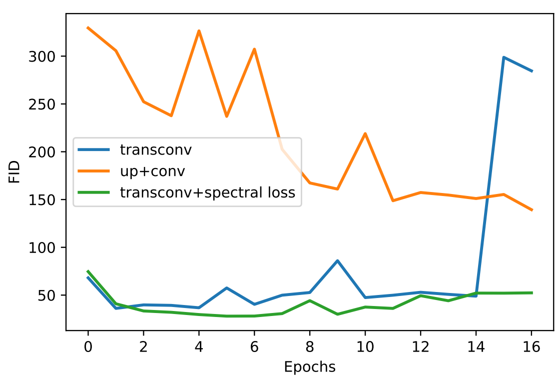


Figure 13: FID (lower is better) over training time for DC-GAN baselines with and without spectral loss (here $\lambda = 2$). While the *up+conv* variant of DCGAN is failing to improve, the FID score over the training time in the *transconv* version is converging but unstable. Only our spectral loss variant is able to achieve low and stable FID scores.

Figures 13 and 14 show the FID evolution along the training epochs, using a baseline GAN implementation with

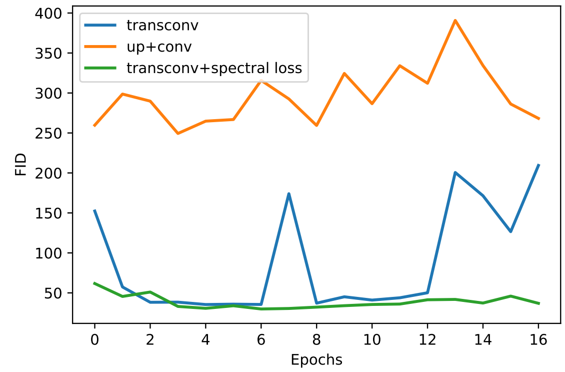


Figure 14: FID (lower is better) over training time for LS-GAN baselines with and without spectral loss (here $\lambda = 0.5$). As for DCGANS, is the *up+conv* variant of LSGAN failing to improve the FID score over the training time. The *transconv* version is converging but unstable. Again, only our spectral loss achieves low and stable FID scores.

different up-convolution units and a corresponding version with spectral loss. These results show an obvious positive effect in terms of the FID measure, where spectral regularization keeps a stable and low FID through out the training while unregularized GANs tend to "collapse". Figure 12 visualizes the correlation between high FID values and failing GAN image generations.

5. Discussion and Conclusion

We showed that common "state of the art" convolutional generative networks, like popular GAN image generators fail to approximate the spectral distributions of real data. This finding has strong practical implications: not only can this be used to easily identify generated samples, it also implies that all approaches towards training data generation or transfer learning are fundamentally flawed and it can not be expected that current methods will be able to approximate real data distributions correctly. However, we showed that there are simple methods to fix this problem: by adding our proposed spectral regularization to the generator loss function and increasing the filter sizes of the final generator convolutions to at least 5×5, we were able to compensate the spectral errors. Experimentally, we have found strong indications that the spectral regularization has a very positive effect on the training stability of GANs. While this phenomenon needs further theoretical investigation, intuitively this makes sense as it is known that high frequent noise can have strong effects on CNN based discriminator networks, which might cause overfitting of the generator.

Source code available:
`https://github.com/cc-hpc-itwm/UpConv`

References

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

[2] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[3] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[4] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

[5] S. Bartunov and D. Vetrov. Few-shot generative modelling with generative matching networks. In *International Conference on Artificial Intelligence and Statistics*, pages 670–678, 2018.

[6] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[7] M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel, A. Dafoe, P. Scharre, T. Zeitzoff, B. Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

[8] R. Chesney and D. Citron. Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff.*, 98:147, 2019.

[9] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and 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*, pages 8789–8797, 2018.

[10] L. Clouâtre and M. Demers. Figr: Few-shot image generation with reptile. *arXiv preprint arXiv:1901.02199*, 2019.

[11] B. Dai, S. Fidler, R. Urtasun, and D. Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] J. Donahue, P. Krähenbühl, and T. Darrell. Adversarial feature learning. *arXiv preprint arXiv:1605.09782*, 2016.

[14] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.

[15] R. Durall, F.-J. Pfreundt, and J. Keuper. Semi few-shot attribute translation. *arXiv preprint arXiv:1910.03240*, 2019.

[16] R. Durall, F.-J. Pfreundt, and J. Keuper. Stabilizing gans with octave convolutions. *arXiv preprint arXiv:1905.12534*, 2019.

[17] A. Dziedzic, J. Paparrizos, S. Krishnan, A. Elmore, and M. Franklin. Band-limited training and inference for convolutional neural networks. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1745–1754, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[19] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[20] 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.

[21] S. Gurumurthy, R. Kiran Sarvadevabhatla, and R. Venkatesh Babu. Deligan: Generative adversarial networks for diverse and limited data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 166–174, 2017.

[22] D. Harris. Deepfakes: False pornography is here and the law cannot protect you. *Duke L. & Tech. Rev.*, 17:99, 2018.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[24] C.-C. Hsu, C.-Y. Lee, and Y.-X. Zhuang. Learning to detect fake face images in the wild. In *2018 International Symposium on Computer, Consumer and Control (IS3C)*, pages 388–391. IEEE, 2018.

[25] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[26] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):107, 2017.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[28] A. K. Jain. *Fundamentals of digital image processing*. Englewood Cliffs, NJ: Prentice Hall,, 1989.

[29] 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.

[30] 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.

[31] Y. Katznelson. *An introduction to harmonic analysis*. Cambridge University Press, 2004.

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

[33] Y. Li, S. Liu, J. Yang, and M.-H. Yang. Generative face completion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3911–3919, 2017.

[34] 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.

[35] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. *arXiv preprint arXiv:1611.08408*, 2016.

[36] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[37] 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.

[38] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva. Detection of gan-generated fake images over social networks. In *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 384–389. IEEE, 2018.

[39] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do gans leave artificial fingerprints? *CoRR*, abs/1812.11842, 2018.

[40] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein. Unrolled generative adversarial networks, 2016.

[41] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[42] S. Mo, M. Cho, and J. Shin. Instance-aware image-to-image translation. In *International Conference on Learning Representations*, 2019.

[43] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395, 2016.

[44] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[45] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[46] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin. Variational autoencoder for deep learning of images, labels and captions. In *Advances in neural information processing systems*, pages 2352–2360, 2016.

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

[48] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[49] 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.

[50] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann. In *NIPS 2017*, 05 2017.

[51] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[52] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[53] Y. Xue, T. Xu, H. Zhang, L. R. Long, and X. Huang. Segan: Adversarial network with multi-scale l 1 loss for medical image segmentation. *Neuroinformatics*, 16(3-4):383–392, 2018.

[54] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5485–5493, 2017.

[55] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. A fourier perspective on model robustness in computer vision. *CoRR*, abs/1906.08988, 2019.

[56] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5505–5514, 2018.

[57] N. Yu, L. Davis, and M. Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *International Conference on Computer Vision (ICCV)*, October 2019.

[58] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.

[59] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1947–1962, 2018.

[60] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[61] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.