Internalized Biases in Fréchet Inception Distance

Anonymous Author(s)
Affiliation
Address
email

Abstract

Fréchet inception distance (FID) established itself as standard performance measuring method for generative adversarial networks (GANs). In this paper, we empirically investigate the biases that are inherited by its underlying design decision of extracting image features using the Inception v3 image classification network. As a result, we investigate how reliable FID is in terms of ranking performances of GANs. In this context, we find that FID is not aligned with human perception and exchanging Inception v3 with different image classification networks simply steers the ranking towards different biases.

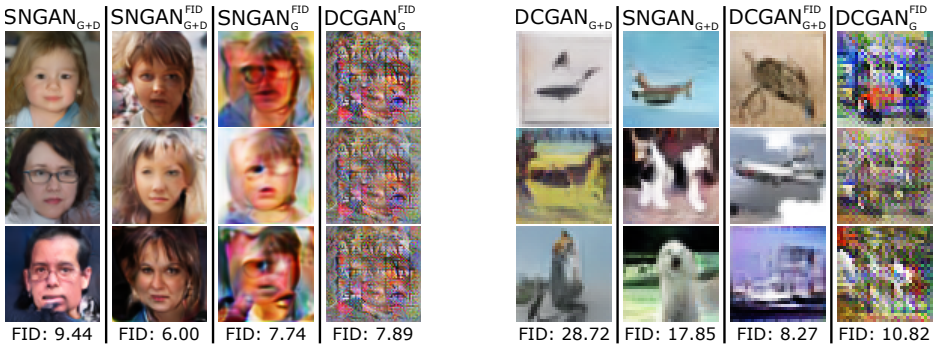


Figure 1: FID [9] is commonly used to decide if one model is superior to another in terms of producing images that are close to the training distribution. Here, we show images generated by different models trained on FFHQ [14] (left) and CIFAR10 [16] (right) with their respective FIDs (lower is better). Would you agree with these rankings?

The generation of photo-realistic, unseen images has made significant progress with the introduction of generative adversarial networks (GANs) [6]. Since then, many architectures emerged competing with each other to provide the best performance [3, 15, 13]. With it arose the question how their performance should be interpreted and measured to provide a ranking between different architectures. Metrics like inception score (IS) and Fréchet inception distance (FID) were proposed to evaluate image distributions automatically. These metrics are based on extracting features from images that are provided by the Inception v3 image classification network [24], and hence inspired their names. While IS was proven to be not useful to compare different models [2], its successor FID is now widely adopted within the GAN community. FID compares the distribution of features between two image datasets that are estimated from training data and samples from the generator network. Well-performing generators are supposed to produce images that match the feature distribution of

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

the data. Some theoretical and practical shortcomings of FID, like a bias due to sample sizes [4] and inconsistent downsampling implementations between different image processing libraries [19] are already discussed in the literature. However, none of these works discuss the shortcomings that come with its underlying feature extractor. The contributions of this paper are as follows:

- We empirically show biases in the underlying assumptions of the commonly used metric Fréchet Inception Distance (FID) caused by the network Inception v3 (section 1).
- We investigate if an improvement in terms of FID is a reliable indicator of improvement in generator performance. Here, we show cases where FID clearly fails to align its ranking with human perception (section 2; see Figure 1 for an example).

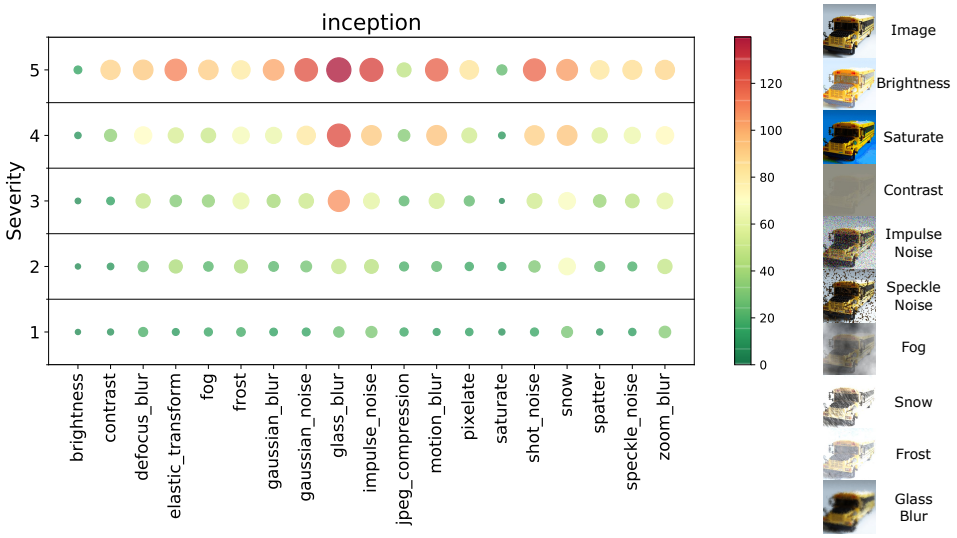


Figure 2: (left) Color-coded FIDs between 19 corruptions of ImageNet validation images with 5 severity levels [8] to their originals. Colors and circle sizes depend on the observed FID over all corruptions and severities. (right) Examples of different corruptions at severity 5.

1 Fréchet Inception Distance and its Biases

To compute FID between two image datasets, image features are extracted by sampling images from those datasets and feeding them into the pretrained Inception v3 [24] image classification network. In the following, we denote Inception v3 by function ϕ that returns feature column vectors given single images and feature matrices with samples on rows given image datasets. FID is then computed via the Wasserstein-2 distance

$$\text{FID} = ||\mu_1 - \mu_2||_2^2 + tr(\Sigma_1) + tr(\Sigma_2) - 2 \cdot tr(\sqrt{\Sigma_1 \Sigma_2}) \tag{1}$$

between two image sources, where $tr(\cdot)$ is the trace operator,

$$\mu_x = \frac{1}{|\mathbf{X}|} \sum_{x \in \mathbf{X}} \phi(x) \qquad \text{and} \qquad \Sigma_x = \frac{(\phi(\mathbf{X}) - \mu_x)^T (\phi(\mathbf{X}) - \mu_x)}{n-1} \tag{2}$$

are the mean vector and covariance matrix of feature vectors of dataset $\mathbf{X}$. The motivation of FID is that, given a sufficiently large number of samples, the first two moments of the feature distributions should match if the images are sampled from the same dataset. Hence, the closer the generator network gets to reproducing the Inception v3 feature distribution of the training data, the smaller FID becomes.

FID is related to model robustness. While pretrained feature extractors can facilitate certain tasks in meaningful ways, like object recognition or detection, we argue that caution needs to be taken in the context of comparing image distributions. The Inception v3 network is trained on the ImageNet object recognition challenge [5], where the task is to classify images into 1000 distinct

classes. Hence, the network learns to extract features from images that discriminate classes found in ImageNet. For this task, it is beneficial for the network to become robust against several distribution shifts, like color or intensity changes and diverse spatial transformations. A common way to equip image classification networks with robustness against before-mentioned shifts is to augment the training data. For example, Inception v3 was trained with the following augmentation pipeline [1]: a) cropping the input image with a random scale (8%-100%) and aspect ratio ($3/4$ to $4/3$), b) randomly flipping the input image horizontally, and c) randomly introducing color distortions in terms of hue, saturation, brightness, and contrast. Consequentially, we assume that the network is at least robust to some degree of corruption. In the following, we investigate i) how robust Inception v3 is against certain corruption types, and ii) how these findings impact the applicability of FID as a metric.

We first look into how different kinds of corruptions influence FID. Research in the area of robustness has produced several benchmark dataset collections, one of which is ImageNet-C [8]. This collection contains the ImageNet [5] validation images with 19 corruption types at 5 severity levels, hence, a total of 95 image datasets. On this benchmark, we can uncover which types of corruptions influence the feature distribution provided by Inception v3 (and therefore FID) to which degree. In Figure 2 we depict the corresponding FIDs between the original ImageNet validation images and all datasets in ImageNet-C. We interpret larger FIDs as indicator that the distribution of features extracted by Inception v3 is influenced more by certain corruptions, and hence we derive that the model is more sensitive to these types of corruptions. We can make the following observations from Figure 2: First, Inception v3 is mostly robust to deviations in brightness, saturation, and contrast up to a certain degree. While information about colors and intensities is lost in these cases, edges are mostly preserved. In the frequency domain, we can see these corruptions acting on low frequencies [25], while high frequencies are mostly untouched. Second, Inception v3 is sensitive to noise that acts on the high frequency spectrum, either by adding high frequency artifacts (impulse noise, shot noise, speckle noise), or by removing high frequency information (glass blur, Gaussian blur). These observations lead us to the assumption that Inception v3 has a bias towards extracting features based on edges and textures rather than color and intensity information. This aligns with its augmentation pipeline that introduces color distortions, but keeps high frequency information intact (in contrast to, for example, augmentation with Gaussian blur). Consequently, FID inherits this bias. **When used as ranking metric, architectures reproducing textures well might be preferred over architectures that reproduce color distributions well**.

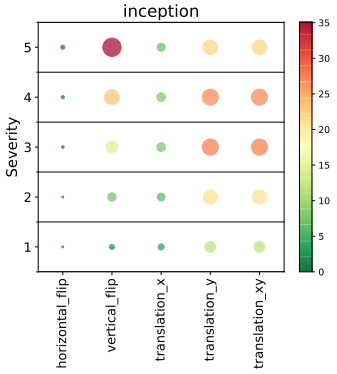


Figure 3: Corruptions introduced by transforming CIFAR10 [16] test images a) with increasing likelihood of horizontal or vertical flips and b) by moving the image in x, y, or both directions up to an increasing distance (reflection padding). The figure depicts the FID between a corruption at a certain level with the original CIFAR10 test dataset.

Since ImageNet-C contains no corruptions that test robustness to translations, we further investigate this aspect by introducing additional corruptions (see Figure 3 in appendix). We are interested in knowing how FID is influenced by a) flipping images either horizontally or vertically, and b) by translating the images in certain directions. Our results indicate that Inception v3 is mostly robust towards horizontal flips and translations, while being more sensitive in the vertical direction. Again, this fits well with the training data augmentation of Inception v3 that reinforces robustness in horizontal directions via random horizontal flips [1].

GAN performance is related to deep fake detection. Before IS and FID were introduced to automatically assess the performance of GANs, user studies were the main source of performance evaluation [27, 21]. Here, the user acts as deep fake detector, and model performance is interpreted as the ability of the model to fool human judgement.

2 Reinforcing the Bias by minimizing FID

If FID is a metric that aligns with human judgement, we can assume that the visual appearance of images generated by any GAN model should improve if we optimize the generator by minimizing FID [17]. But can this assumption hold? To verify, we train two GAN architectures, DCGAN [20]

and SNGAN [18], on FFHQ [14] downscaled to 64^2 image resolution (further called FFHQ64) and CIFAR10 [16]. We consider three training procedures, a) $\text{GAN}_{\text{G+D}}$ where we train the model in its common setting in which a discriminator network provides the training signal for the generator network (baseline), b) $\text{GAN}_{\text{G+D}}^{\text{FID}}$ where we extend a) by adding an additional loss for the generator network given by Equation 1 (further called FID loss), and c) $\text{GAN}_{\text{G}}^{\text{FID}}$ where we drop the discriminator network and train the generator solely by minimizing the FID loss. Depictions of b) and c) are given in Figure 6. In each iteration that we minimize FID loss directly, we generate 400 images for DCGAN and 360 images for SNGAN to approximate the current FID. During training we measure FID to track the training progress after each epoch by sampling $10\,000$ images from the generator.

The progress for each training setting is shown in Figure 4. Given the training curves provided by evaluating FID after each epoch, one could assume that applying FID loss stabilizes the training. We could go even further and assume that the discriminator is not even necessary for GAN training, since the results measured in FID differ only slightly. However, the story changes after examining the images each of these models generate (see Figure 1 and appendix for further examples). By inspecting samples, we observe no visual improvement in image quality between $\text{GAN}_{\text{G+D}}$ and $\text{GAN}_{\text{G+D}}^{\text{FID}}$ despite the large gaps in terms of FID. We even argue that images produced by $\text{GAN}_{\text{G+D}}^{\text{FID}}$ contain more artifacts and are less visually pleasing, which we see as a sign of overfitting to the FID loss. We hypothesize that the generator learns to produce unsuitable features to match the training data distribution. This observation becomes more severe in the case of $\text{GAN}_{\text{G}}^{\text{FID}}$. Here, we notice that the missing discriminator leads to spatially incoherent feature distributions. For example $\text{SNGAN}_{\text{G}}^{\text{FID}}$ adds mostly single eyes and aligns facial characteristics in a daunting manner. We assume that this is the result of the choice of layer that features are extracted from Inception v3. Features are spatially pooled and therefore lose all spatial information before they are extracted. While human annotators would surely prefer images produced by $\text{SNGAN}_{\text{D+G}}$ over $\text{SNGAN}_{\text{G}}^{\text{FID}}$ (in cases where data fidelity is preferred over art), we see that this is not reflected by FID. **Hence, FID is not aligned with human perception**. We argue that discriminative features provided by image classification networks are not sufficient to provide the basis of a meaningful metric. To further emphasize this point, we substitute Inception v3 by an extensive choice of different image classification networks. We evaluate FIDs on ImageNet-C at different severities for each backbone network (details are provided in the appendix). Here we see that biases present in Inception v3 are also widely present in other classification networks. Additionally, we see that different networks would produce different rankings in-between corruption types.

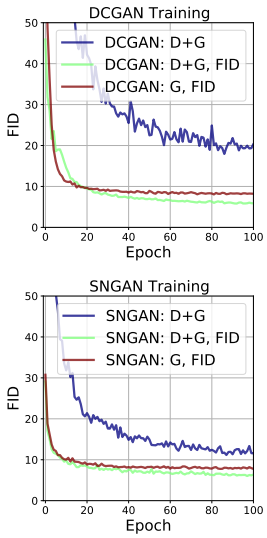


Figure 4: Results of training DCGAN and SNGAN with different training procedures. FID loss stabilizes the training and achieves better performances (smaller FID), whereas the discriminator only marginally improves the generator performance.

3 Conclusion

We provided an overview of inherent biases that are present due to the design decision of FID to use Inception v3 as feature extractor. We showed that the way Inception v3 was trained enforced it to become robust to corruptions related to color, intensity, saturation, and horizontal translations, while being sensitive to corruptions of textures and to vertical translations. These biases influence the ranking when different architectures are compared and need to be taken into account. Additionally, we showed that FID as a metric is not aligned with human perception by minimizing FID as a loss term. Here, we showed that generators trained with FID loss produce images with substantially improved FIDs, but worse visual appearance. The plotted training curves showing the development of FID are misleading in those cases. Finally, we showed that substituting Inception v3 with another image classification network would simply mean to exchange different biases. Hence, we hope to inspire further research to close the gap towards a humanly aligned and unbiased metric that enables to fairly rank generator architectures.

References

[1] Advanced Guide to Inception v3 on Cloud TPU. `https://cloud.google.com/tpu/docs/inception-v3-advanced`. Accessed: 2021-10-08.

[2] Shane T. Barratt and Rishi Sharma. A Note on the Inception Score. *ArXiv*, abs/1801.01973, 2018.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.

[4] Min Jin Chong and D. Forsyth. Effectively Unbiased FID and Inception Score and Where to Find Them. In *CVPR*, 2020.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[6] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative Adversarial Nets. In *NIPS*, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR*, 2019.

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and S. Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NIPS*, 2017.

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[12] Shyo Prakash Jakhar, Amita Nandal, and Rahul Dixit. Classification and measuring accuracy of lenses using inception model v3. In Manoj Kumar Sharma, Vijaypal Singh Dhaka, Thinagaran Perumal, Nilanjan Dey, and João Manuel R. S. Tavares, editors, *Innovations in Computational Intelligence and Computer Vision*, pages 376–383, Singapore, 2021. Springer Singapore.

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[14] 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.

[15] 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*, pages 8110–8119, 2020.

[16] A. Krizhevsky. Learning Multiple Layers of Features from Tiny Images. 2009.

[17] Alexander Mathiasen and Frederik Hvilshøj. Backpropagating through Fréchet Inception Distance, 2021.

[18] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[19] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Buggy Resizing Libraries and Surprising Subtleties in FID Calculation. *arXiv*, abs/12104.11222, 2021.

[20] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[21] Tim Salimans, I. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *NIPS*, 2016.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[23] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2016.

[25] Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A Fourier Perspective on Model Robustness in Computer Vision. In *NIPS*, 2019.

[26] Xingcheng Zhang, Zhizhong Li, Chen Change Loy, and Dahua Lin. Polynet: A pursuit of structural diversity in very deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 718–726, 2017.

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017.

[28] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.

Supplementary Material

Internalized Biases in Fréchet Inception Distance

Code, trained models and datasets containing generated images will be published on GitHub. In this supplementary, we depict the Inception v3 architecure (A), provide more details about our experiments in section 2 as well as an overview of all results (B-C), show that it is possible to improve generator architectures by minimizing FID (D), show more generated images for each model trained (E-H), compute "FIDs" with different backbone classification networks on ImageNet-C (I), and describe an experiment for deep fake detection using FID (J), .

- Appendix A: Inception v3 architecture
- Appendix B: Training Settings (section 2)
- Appendix C: Implementation Details to Minimizing FID (section 2)
- Appendix D: Improving generator architectures with FID loss
- Appendix E: Generated Images (DCGAN/FFHQ)
- Appendix F: Generated Images (SNGAN/FFHQ)
- Appendix G: Generated Images (DCGAN/CIFAR10)
- Appendix H: Generated Images (SNGAN/CIFAR10)
- Appendix I: FIDs when substituting backbone networks on ImageNet-C
- Appendix J: Deep Fake Detection with FID

Appendix A: Inception v3 architecture

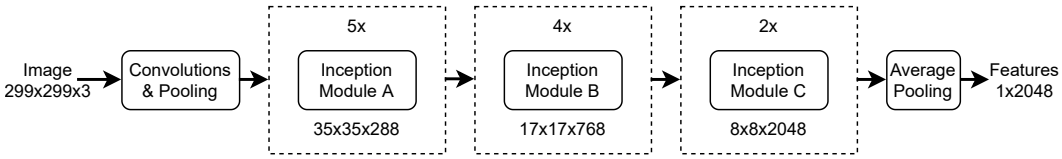


Figure 5: Depiction of the Inception v3 architecture (inspired by [12]) that is used to extract image features to compute FID [9]. Grid size reductions are not shown here.

Appendix B: Training Settings (section 2)

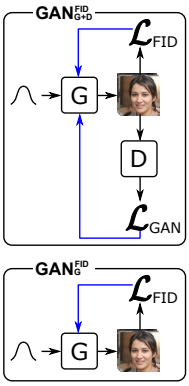


Figure 6: Different settings to train generator networks. In $\text{GAN}^{\text{FID}}_{\text{G+D}}$ we add an additional loss called FID loss (see Equation 1). In $\text{GAN}^{\text{FID}}_{\text{G}}$ we drop the discriminator network and train the generator solely by minimizing FID loss.

Appendix C: Implementation Details to Minimizing FID (section 2)

We train each of the models for 48 hours and report the best FID measured during training. FIDs are measured after each epoch. We train DCGAN with the minimax loss

$$\mathcal{L}_G = \mathbb{E}_{z\sim N(0,1)}[\log(1 - D(G(z)))], \tag{3}$$

$$\mathcal{L}_D = -\mathbb{E}_{x\sim \text{data}}[\log(D(x))] \; - \; \mathbb{E}_{z\sim N(0,1)}[\log\left(1 - D(\hat{\mathbf{x}})\right)], \tag{4}$$

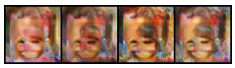
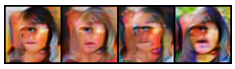
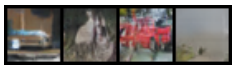
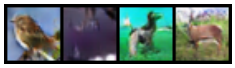
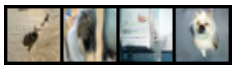
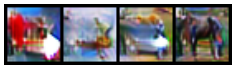
and for optimization we use Adam with $\beta = (0.5, 0.999)$, $\epsilon = 10^{-8}$, learning rate $= 0.0002$, and weight decay $= 0$.
We train SNGAN with the hinge loss

$$\mathcal{L}_G = -\mathbb{E}_{z\sim N(0,1)}[D(G(z))], \tag{5}$$

$$\mathcal{L}_D = -\mathbb{E}_{x\sim \text{data}}[\min(0, -1 + D(x))] - \mathbb{E}_{z\sim N(0,1)}[\min(0, -1 - D(G(z)))]. \tag{6}$$

and for optimization we use Adam with $\beta = (0.0, 0.9)$, $\epsilon = 10^{-8}$, learning rate $= 0.0002$, and weight decay $= 0$.
We combine all results in the following table.

Table 1: Combined results of trained models from section 2.

Data	Resolution	Model	FID↓	Images
FFHQ	64^2	$\text{DCGAN}_{\text{G+D}}$	14.86	
		$\text{DCGAN}^{\text{FID}}_{\text{G+D}}$	5.38	
		$\text{DCGAN}^{\text{FID}}_{\text{G}}$	7.89	
		$\text{DCGAN-Up}^{\text{FID}}_{\text{G}}$	7.61	
		$\text{SNGAN}_{\text{G+D}}$	9.44	
		$\text{SNGAN}^{\text{FID}}_{\text{G+D}}$	6.00	
		$\text{SNGAN}^{\text{FID}}_{\text{G}}$	7.74	
CIFAR10	32^2	$\text{DCGAN}_{\text{G+D}}$	28.72	
		$\text{DCGAN}^{\text{FID}}_{\text{G+D}}$	8.27	
		$\text{DCGAN}^{\text{FID}}_{\text{G}}$	10.82	
		$\text{DCGAN-Up}^{\text{FID}}_{\text{G}}$	10.77	
		$\text{SNGAN}_{\text{G+D}}$	17.85	
		$\text{SNGAN}^{\text{FID}}_{\text{G+D}}$	8.07	
		$\text{SNGAN}^{\text{FID}}_{\text{G}}$	11.66	

Appendix D: Improving generator architectures with FID loss

Images produced by $\text{DCGAN}_{\text{G}}^{\text{FID}}$ (see Figure 1 and Figure 10) are only hardly recognizable due to noise covering the images. In this case, minimizing FID without discriminator exposes possible flaws in its architecture. Indeed, when we exchange transpose convolutions in DCGAN with bilinear upsampling followed by convolutional kernels we observe that the noise vanishes (see Figure 7).

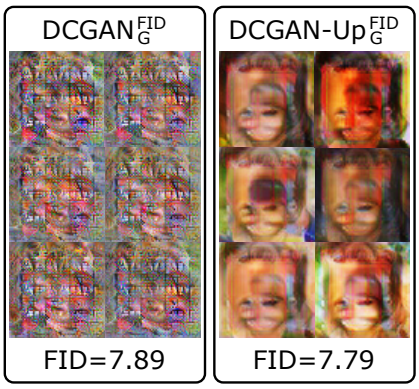


Figure 7: DCGAN TransposeConv vs UpsampleConv FFHQ.

Appendix E: Generated Images (DCGAN/FFHQ)

Figure 8: $\text{DCGAN}_{\text{G+D}}$ trained on FFHQ64. FID: 14.86.

Figure 9: $\text{DCGAN}^{\text{FID}}_{\text{G+D}}$ trained on FFHQ64. FID: 5.38.

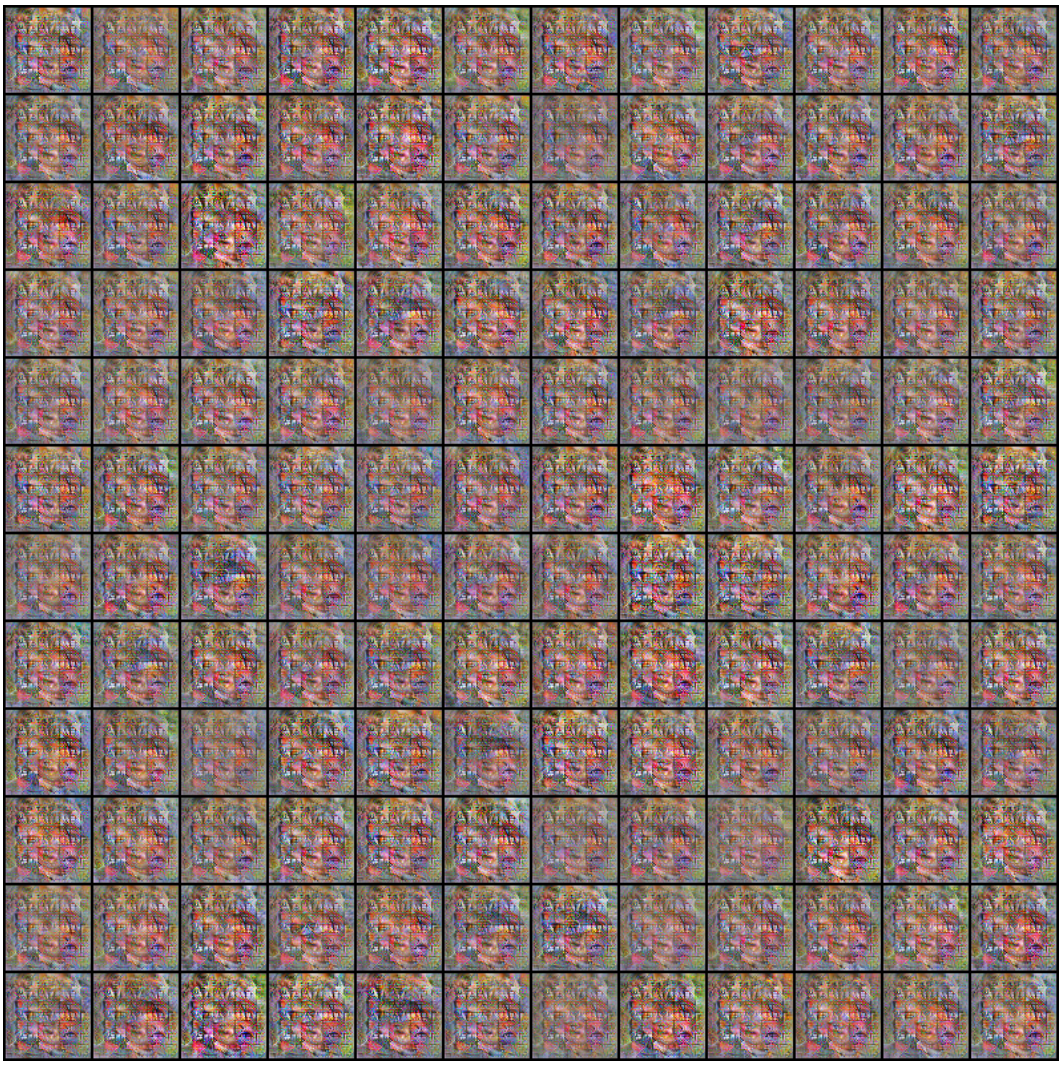


Figure 10: $\text{DCGAN}_{\text{G}}^{\text{FID}}$ trained on FFHQ64. FID: 7.89.

Figure 11: DCGAN-Up$_{\mathrm{G}}^{\mathrm{FID}}$ (upsampling instead of transpose convolutions) trained on FFHQ64. FID: 7.61.

Appendix F: Generated Images (SNGAN/FFHQ)

Figure 12: $\text{SNGAN}_{\text{G+D}}$ trained on FFHQ64. FID: 9.44.

Figure 13: $\text{SNGAN}^{\text{FID}}_{\text{G+D}}$ trained on FFHQ64. FID: 6.00.

Figure 14: $\text{SNGAN}_{\text{G}}^{\text{FID}}$ trained on FFHQ64. FID: 7.74.

Appendix G: Generated Images (DCGAN/CIFAR10)

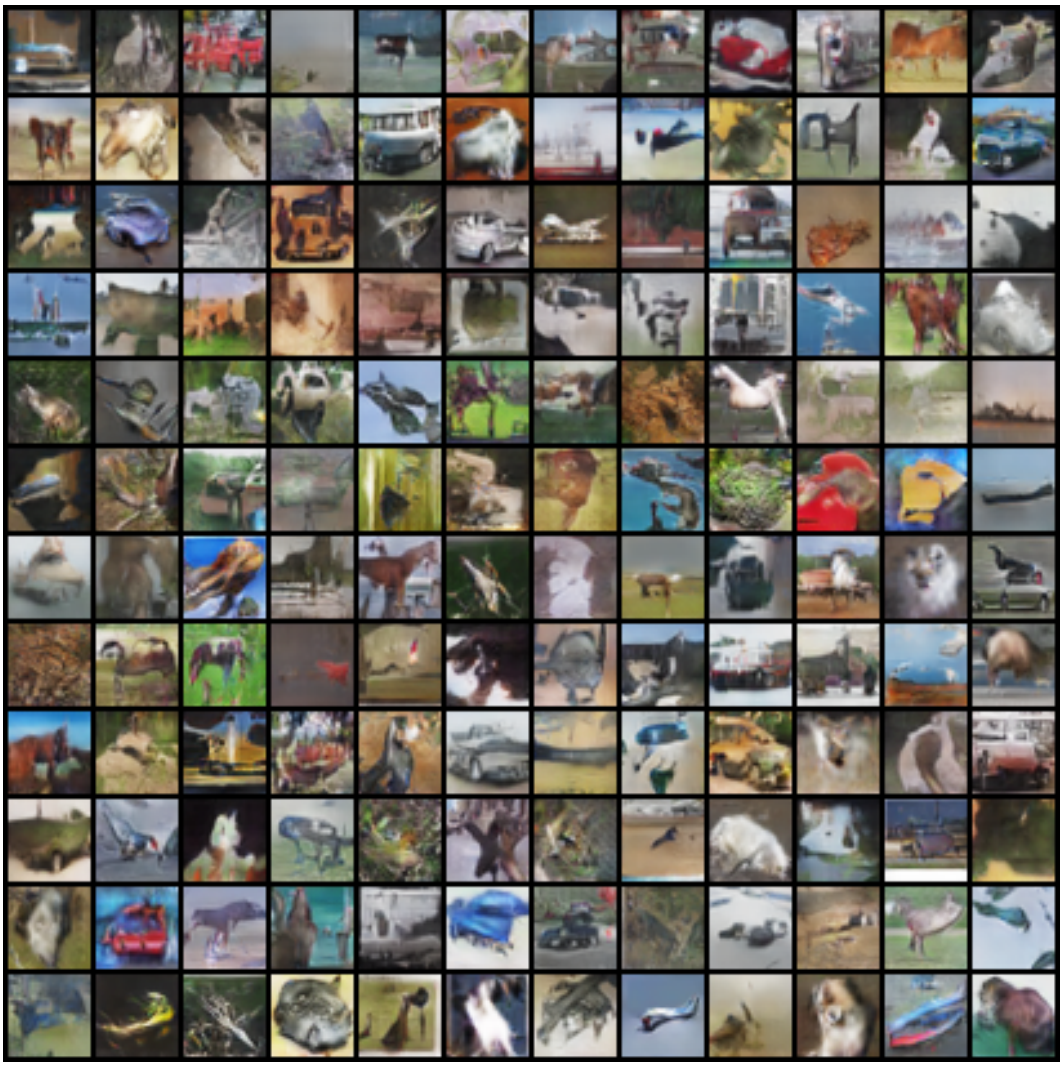


Figure 15: $\text{DCGAN}_{\text{G+D}}$ trained on CIFAR10. FID: 28.72.

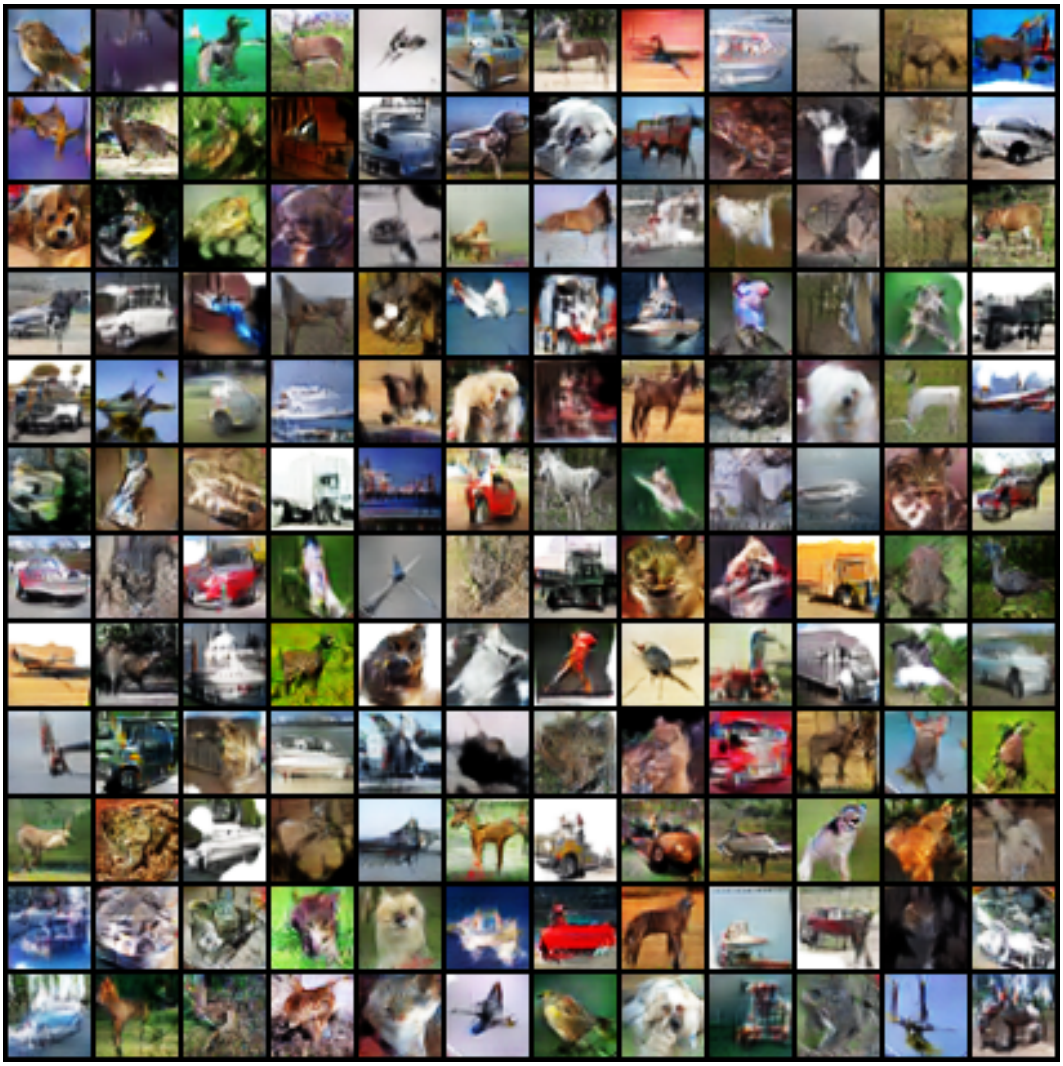


Figure 16: $\text{DCGAN}^{\text{FID}}_{\text{G+D}}$ trained on CIFAR10. FID: 8.27.

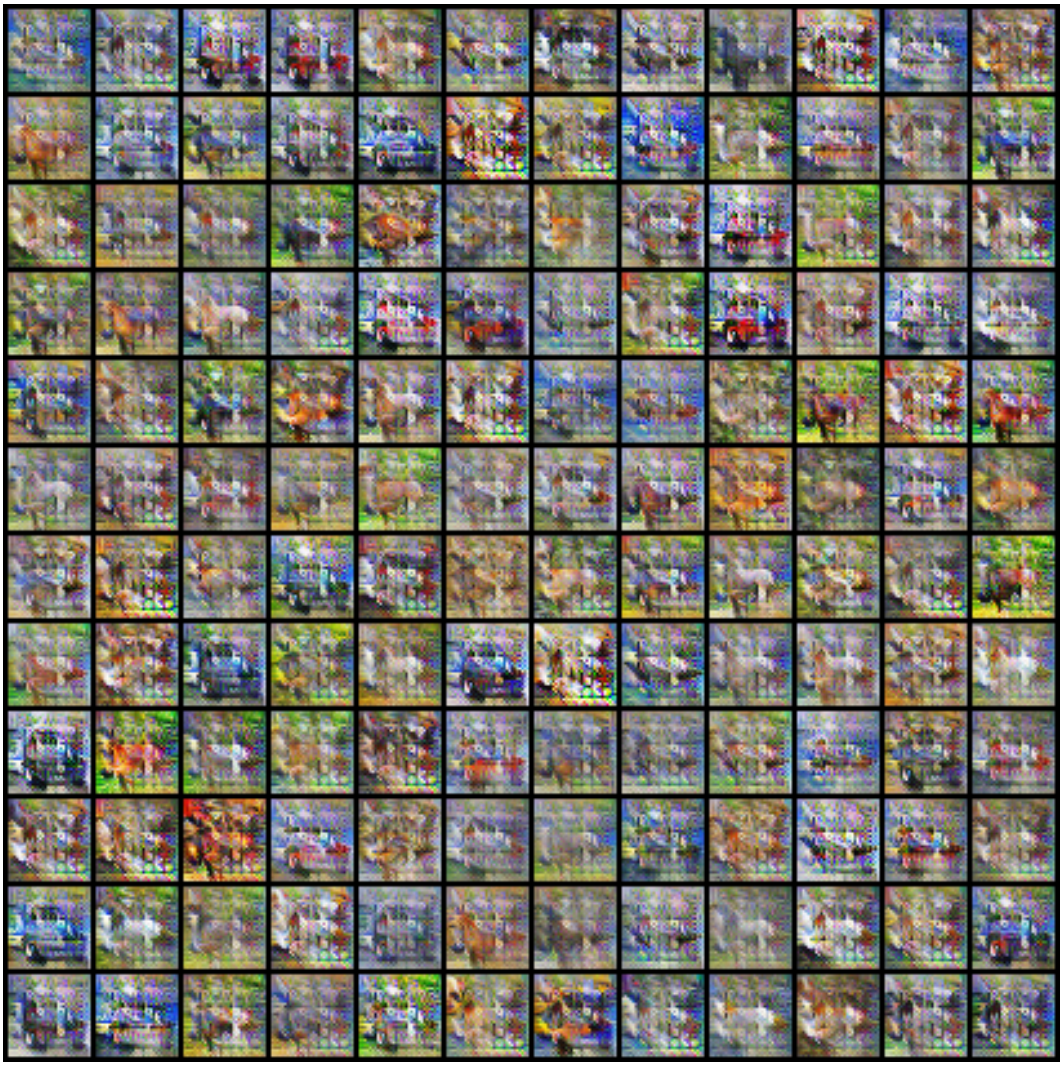


Figure 17: $\text{DCGAN}_{\text{G}}^{\text{FID}}$ trained on CIFAR10. FID: 10.82.

Figure 18: DCGAN-Up$_{\mathrm{G}}^{\mathrm{FID}}$ (upsampling instead of transpose convolutions) trained on CIFAR10. FID: 10.77.

Appendix H: Generated Images (SNGAN/CIFAR10)

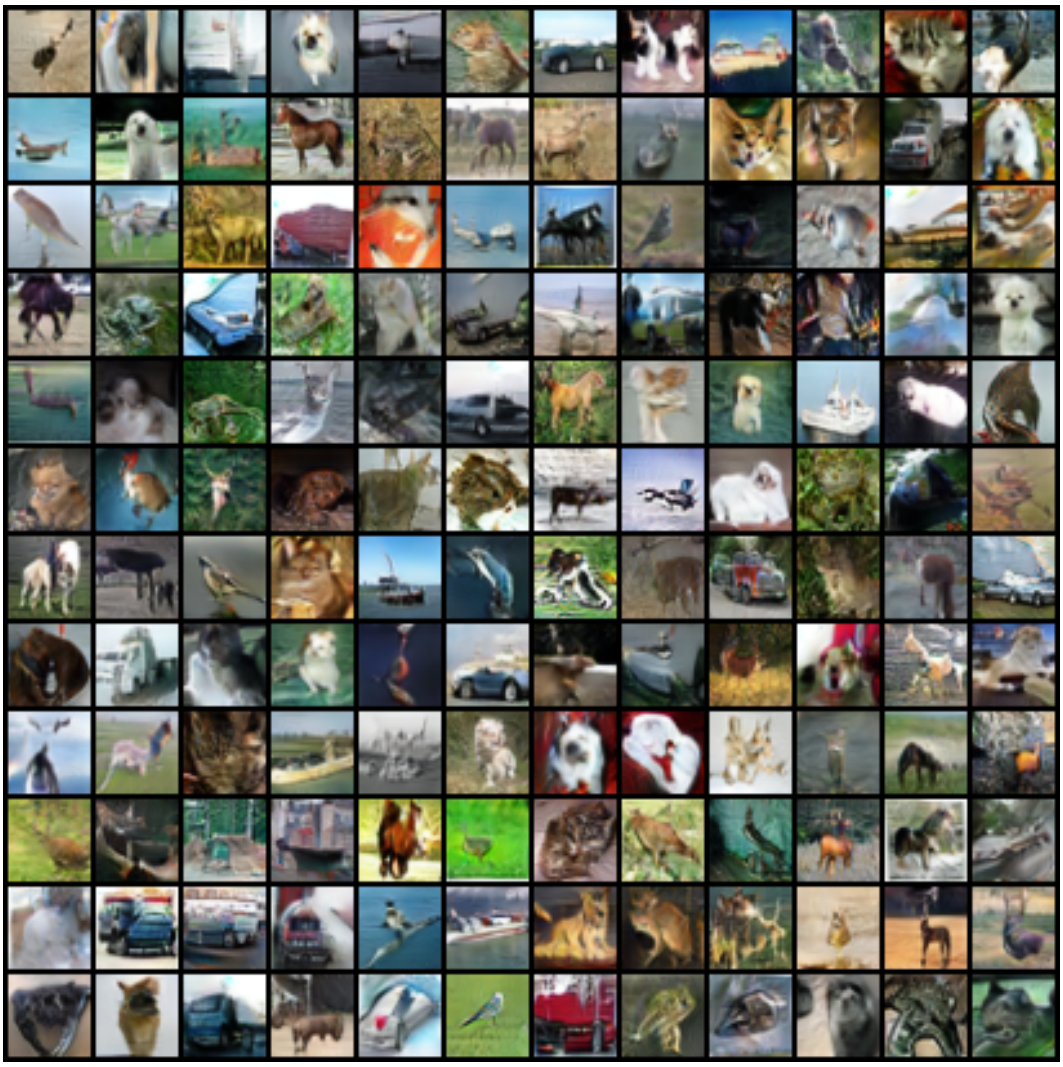


Figure 19: $\text{SNGAN}_{\text{G+D}}$ trained on CIFAR10. FID: 17.85.

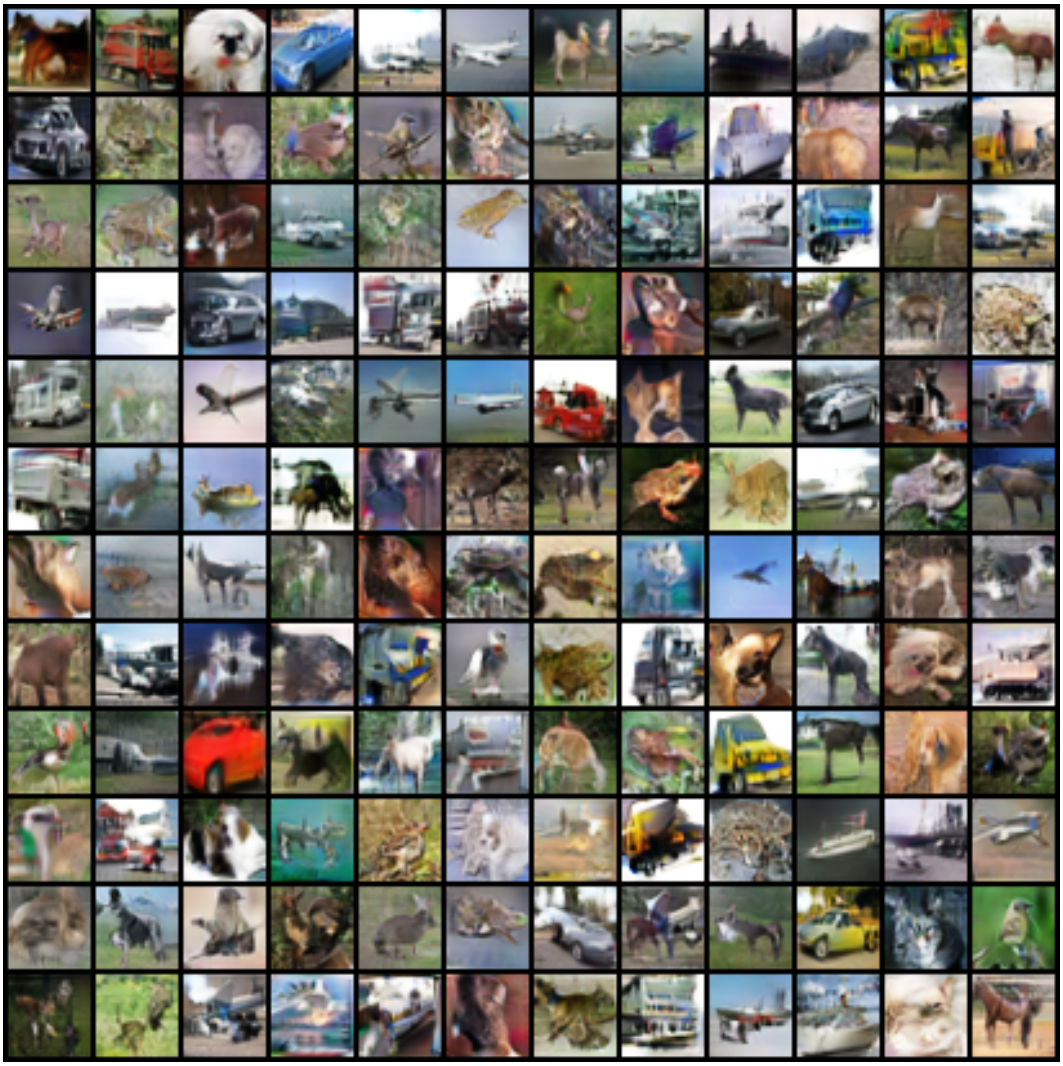


Figure 20: $\text{SNGAN}^{\text{FID}}_{\text{G+D}}$ trained on CIFAR10. FID: 8.07.

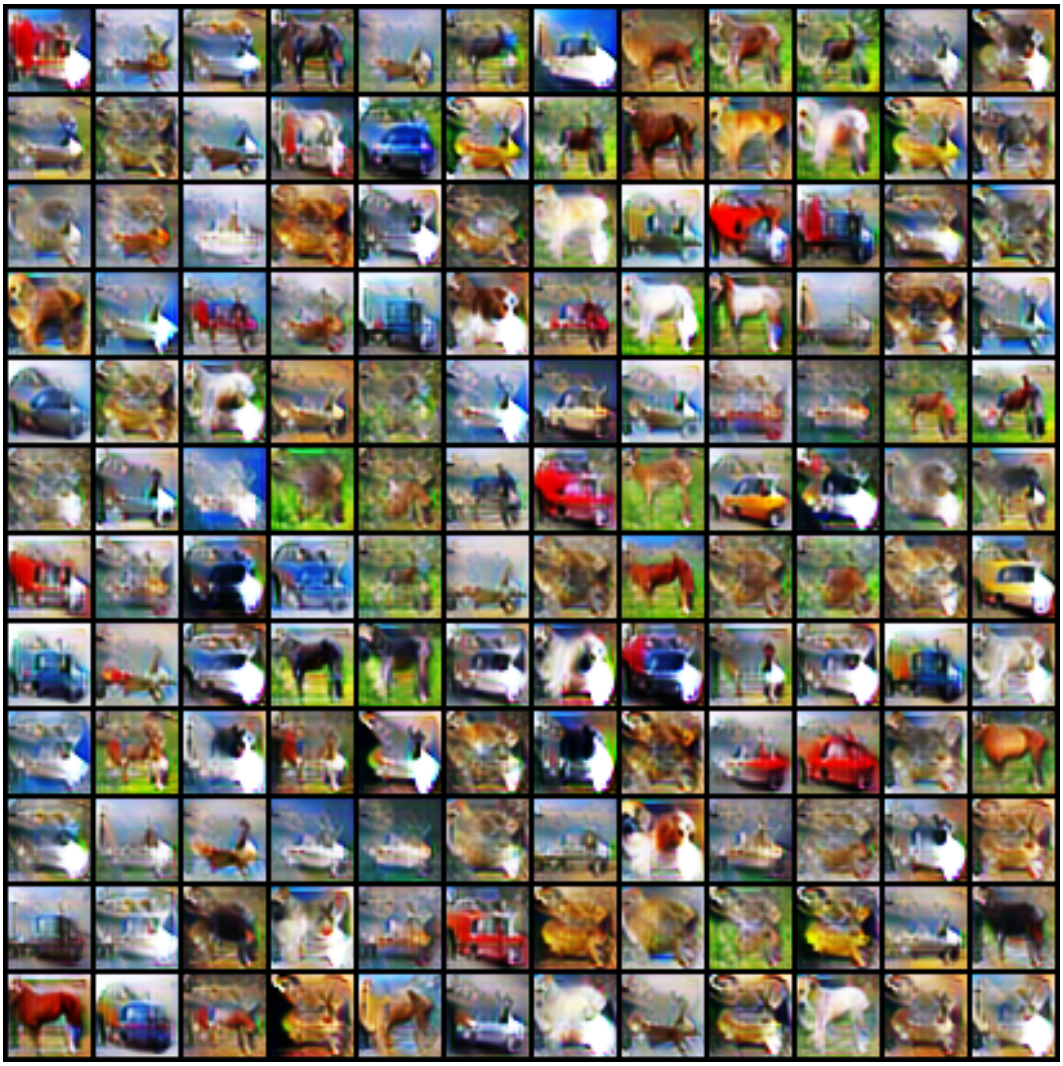


Figure 21: $\text{SNGAN}_{\text{G}}^{\text{FID}}$ trained on CIFAR10. FID: 11.66.

Appendix I: FIDs when substituting backbone networks on ImageNet-C

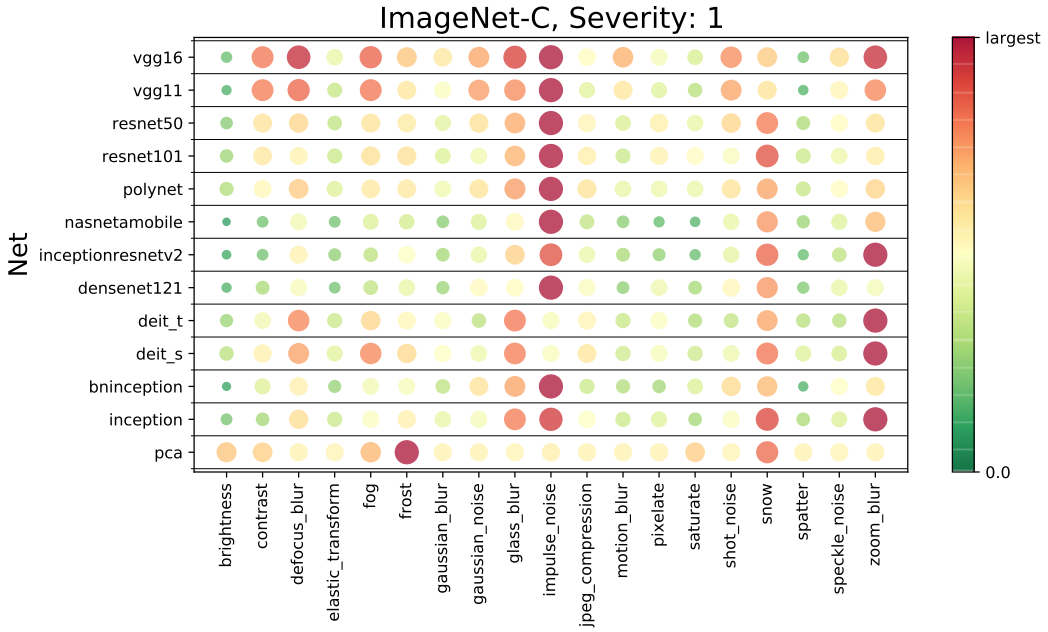


Figure 22: Color-coded FIDs between ImageNet validation images and 19 corrupted versions thereof provided by ImageNet-C. Inception v3 is substituted by different classification networks [22, 7, 26, 28, 23, 10, 11, 24] to see how FID would have been All corruptions are at severity 1. Colors and circle sizes depend on the largest observed FID per network. Additionally, principal component analysis (PCA) is shown, which provides descriptive features with different sensitivity to corruptions compared to image classification networks.

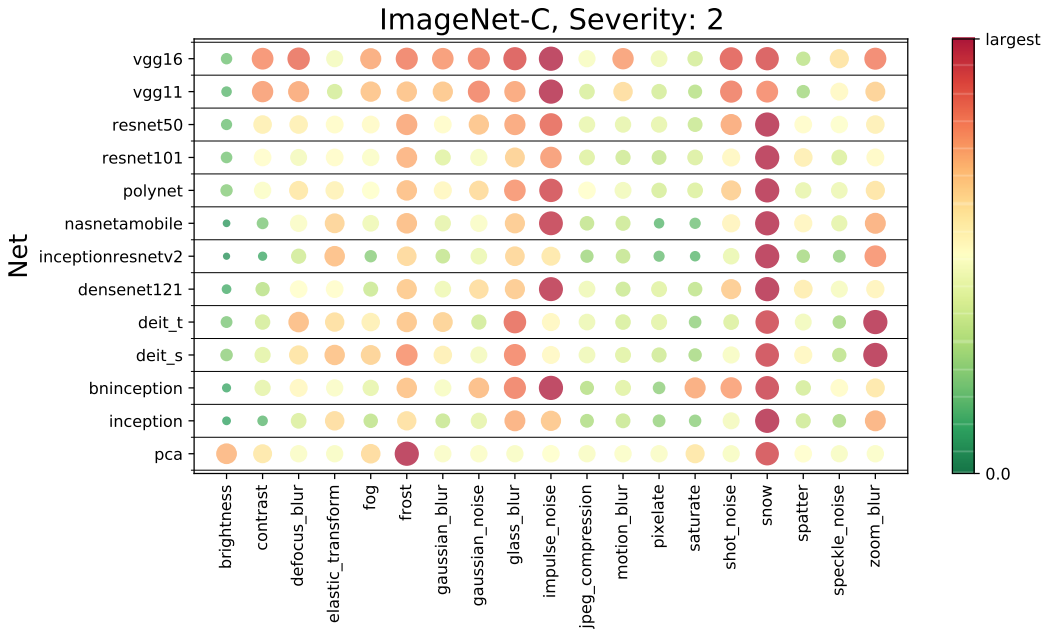


Figure 23: Color-coded FIDs between ImageNet validation images and 19 corrupted versions thereof provided by ImageNet-C. Inception v3 is substituted by different classification networks [22, 7, 26, 28, 23, 10, 11, 24] to see how FID would have been All corruptions are at severity 2. Colors and circle sizes depend on the largest observed FID per network. Additionally, principal component analysis (PCA) is shown, which provides descriptive features with different sensitivity to corruptions compared to image classification networks.

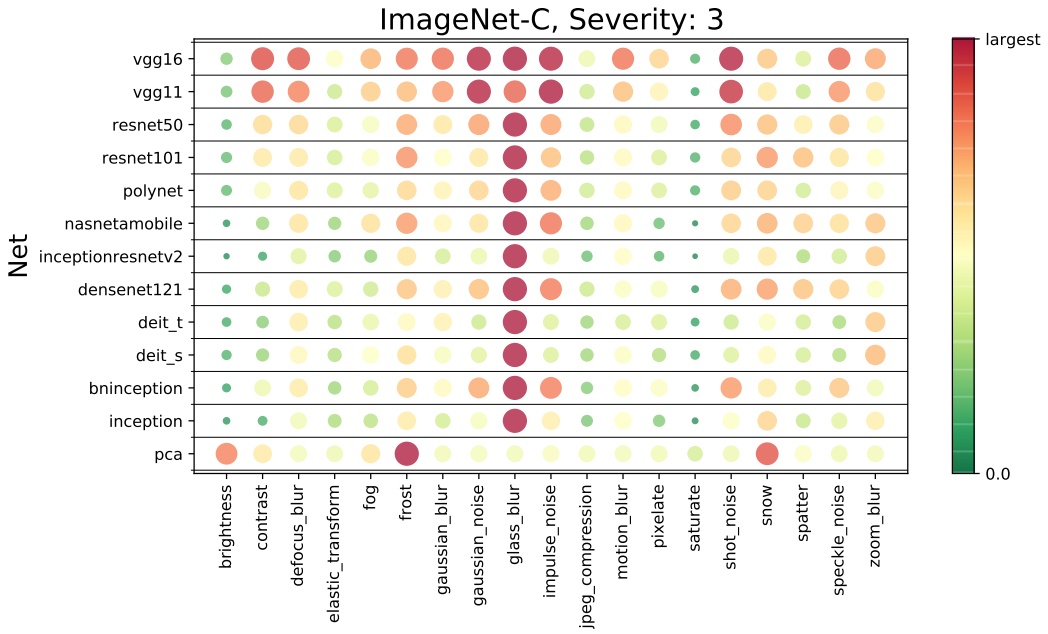


Figure 24: Color-coded FIDs between ImageNet validation images and 19 corrupted versions thereof provided by ImageNet-C. Inception v3 is substituted by different classification networks [22, 7, 26, 28, 23, 10, 11, 24] to see how FID would have been All corruptions are at severity 3. Colors and circle sizes depend on the largest observed FID per network. Additionally, principal component analysis (PCA) is shown, which provides descriptive features with different sensitivity to corruptions compared to image classification networks.

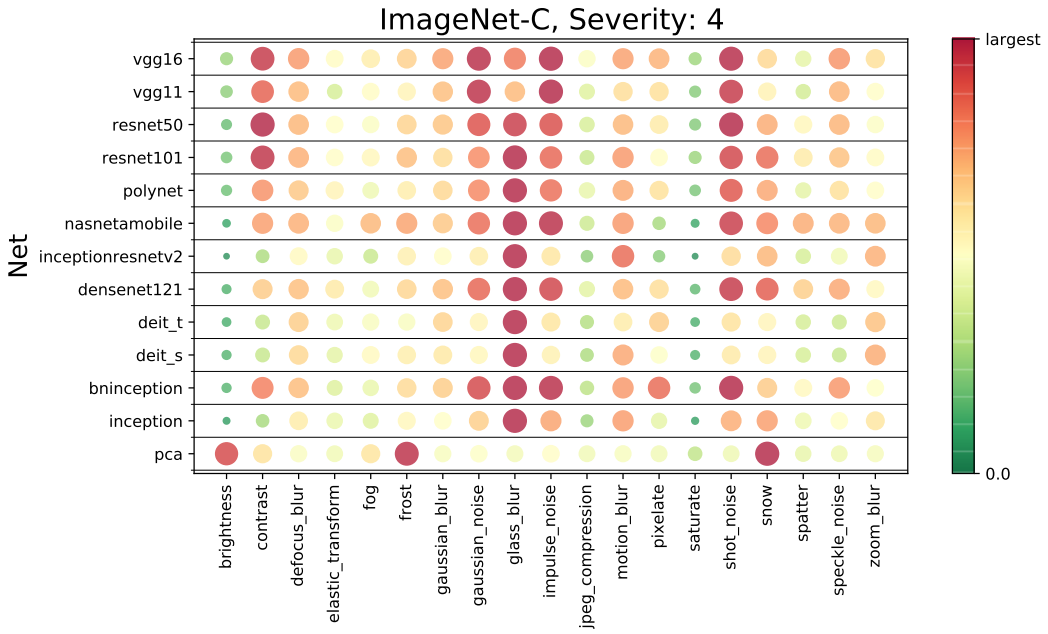


Figure 25: Color-coded FIDs between ImageNet validation images and 19 corrupted versions thereof provided by ImageNet-C. Inception v3 is substituted by different classification networks [22, 7, 26, 28, 23, 10, 11, 24] to see how FID would have been All corruptions are at severity 4. Colors and circle sizes depend on the largest observed FID per network. Additionally, principal component analysis (PCA) is shown, which provides descriptive features with different sensitivity to corruptions compared to image classification networks.

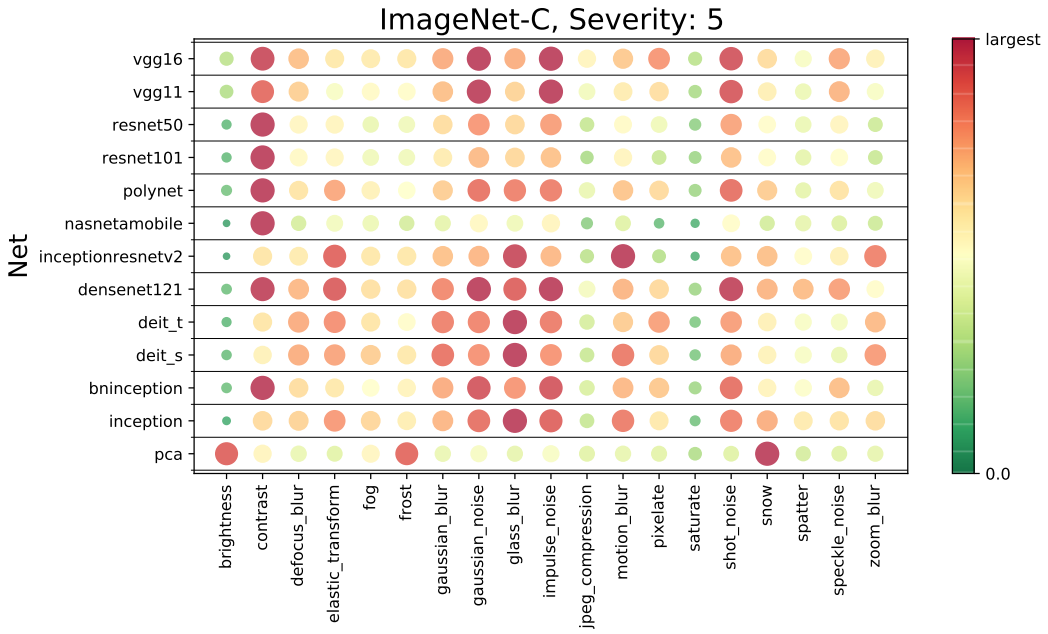


Figure 26: Color-coded FIDs between ImageNet validation images and 19 corrupted versions thereof provided by ImageNet-C. Inception v3 is substituted by different classification networks [22, 7, 26, 28, 23, 10, 11, 24] to see how FID would have been All corruptions are at severity 5. Colors and circle sizes depend on the largest observed FID per network. Additionally, principal component analysis (PCA) is shown, which provides descriptive features with different sensitivity to corruptions compared to image classification networks.

Appendix J: Deep Fake Detection with FID

We map 70 000 images from FFHQ and 70 000 images generated by StyleGAN2 (with and without truncation) into the Inception v3 feature space by using the FID PyTorch implementation. We split each into 60 000 training and 10 000 testing images, and hence, end up with balanced training datasets containing 120 000 images and balanced test datasets containing 20 000 images. We train sklean logistic regression models and report the accuracy on the test dataset. A selection of corresponding images is shown in Figure 27. We argue that a meaningful metric should be visually aligned with human perception. Hence, if a human can be fooled by a generator network, then this generator should be considered superior to one that is not able to do so. We see that truncation decreases FID significantly, and consequently improves the ability of detecting StyleGAN2 generated images as fake. In contrast, we show images produced by StylegAN2 without and with truncation in Figure 27. By inspecting the images we observe that truncation removes textures (and also artifacts). We hypothesize that its bias towards textures facilitates Inception v3 to extract features that allow almost perfect detectability (98% when truncation is applied). However, we expect humans to be more easily fooled by truncated images than untruncated ones. Hence, we argue that this is a hint towards that FID is not aligned with human perception.

Table 2: A logistic regression is trained to perform fake detection based on image features provided by Inception v3.

Model	**Data Res.**	**FID↓**	**Accuracy↑**	**Precision↑**	**Recall↑**	**F1↑**
StyleGAN2 $\psi = 1.0$	FFHQ 1024^2	2.65	.71	.70	.72	.71
StyleGAN2 $\psi = 0.5$	FFHQ 1024^2	57.77	.98	.98	.98	.98

Figure 27: StyleGAN2 images.