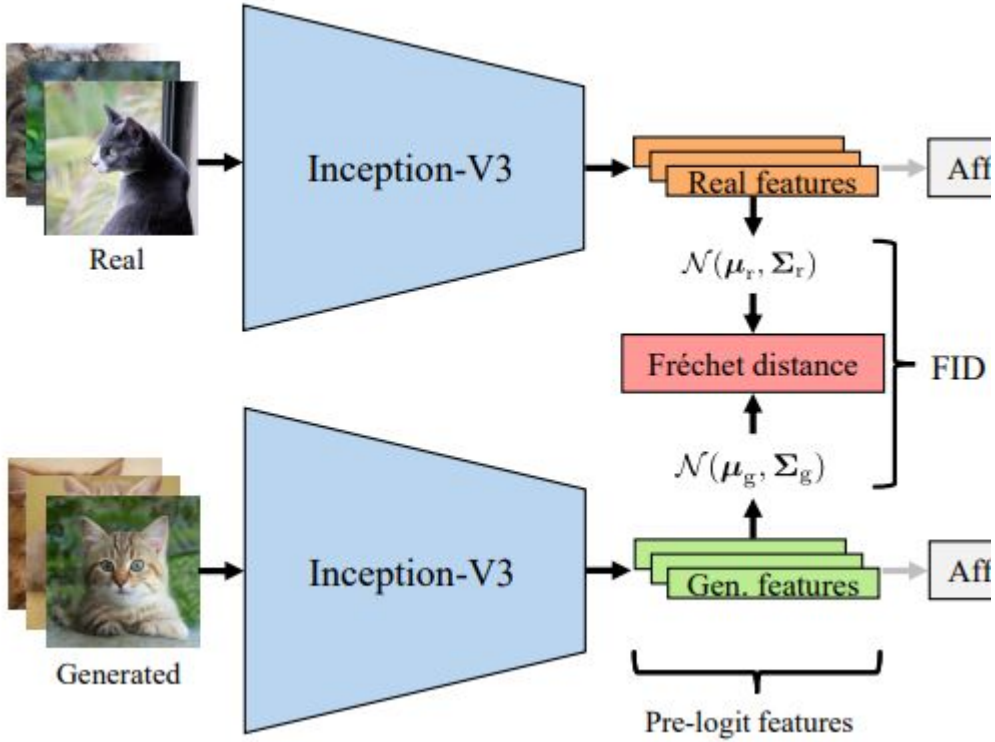
The Role of ImageNet Classes in Fréchet Inception Distance

Sangyun Lee

Fréchet Inception Distance



"perceptual distance between two distribution"

is it true?

$$\mathrm{FID}\left(\boldsymbol{\mu}_r, \boldsymbol{\Sigma}_r, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g\right) = \left\|\boldsymbol{\mu}_r - \boldsymbol{\mu}_g\right\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_g - 2\left(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g\right)^{\frac{1}{2}}\right), \qquad (1)$$

What does FID look at in an image?

$$\boldsymbol{\mu}'_{\mathrm{g}} = \frac{N-1}{N}\boldsymbol{\mu}_{\mathrm{g}} + \frac{1}{N}\boldsymbol{f},$$

$$\boldsymbol{\Sigma}'_{\mathrm{g}} = \frac{N-2}{N-1}\boldsymbol{\Sigma}_{\mathrm{g}} + \frac{1}{N}(\boldsymbol{f} - \boldsymbol{\mu}_{\mathrm{g}})^T(\boldsymbol{f} - \boldsymbol{\mu}_{\mathrm{g}})).$$

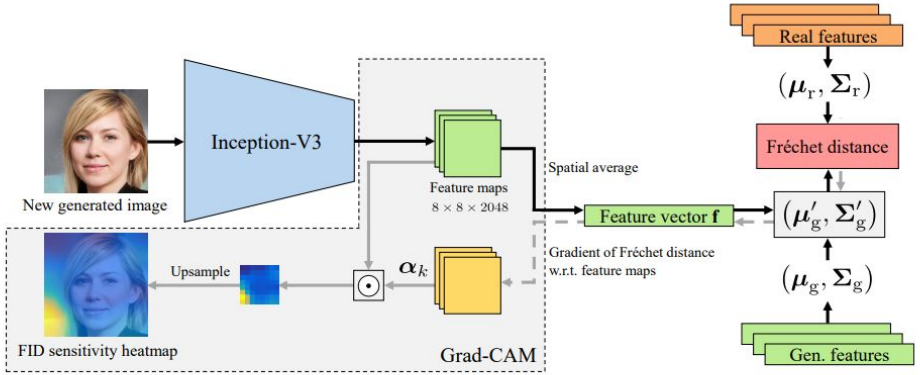


Figure 2: Visualizing which regions of an image FID is the most sensitive to. We augment the pre-computed feature statistics with a newly generated image, compute the FID, and use Grad-CAM [51] to visualize the spatial importance in low-resolution feature maps that are subsequently upsampled to match the input resolution.

Pebay, P.P.: Formulas for robust, one-pass parallel computation of covariances and arbitrary-order statistical moments. Tech. Rep. SAND2008-6212, Sandia National Laboratories (2008)

What does FID look at in an image?

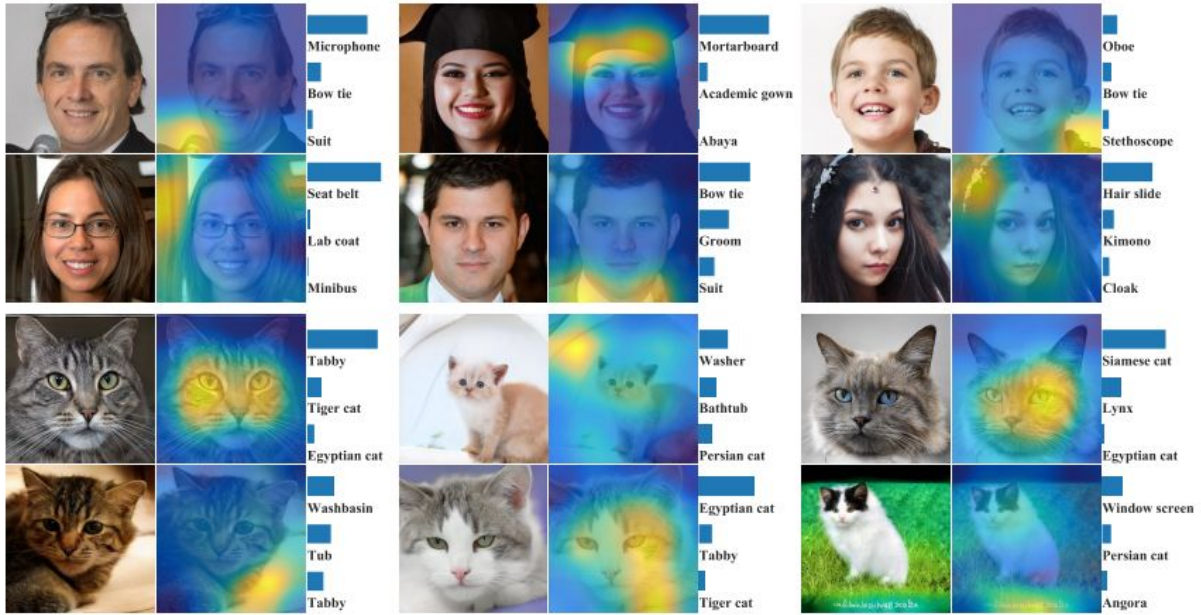


Figure 3: StyleGAN2-generated images along with heatmap visualizations of the image regions that FID considers important in FFHQ (top) and LSUN CAT (bottom). Yellow indicates regions that are more important and blue regions that are less important, i.e., modifying the content of the yellow regions affects FID most strongly. As many of the yellow areas are completely outside the intended subject, we sought an explanation from the Top-3 ImageNet class predictions for the generated image. It turns out that FID is very strongly focused on the area that corresponds to the predicted Top-1 class — whatever that may be.

What does FID look at in an image?

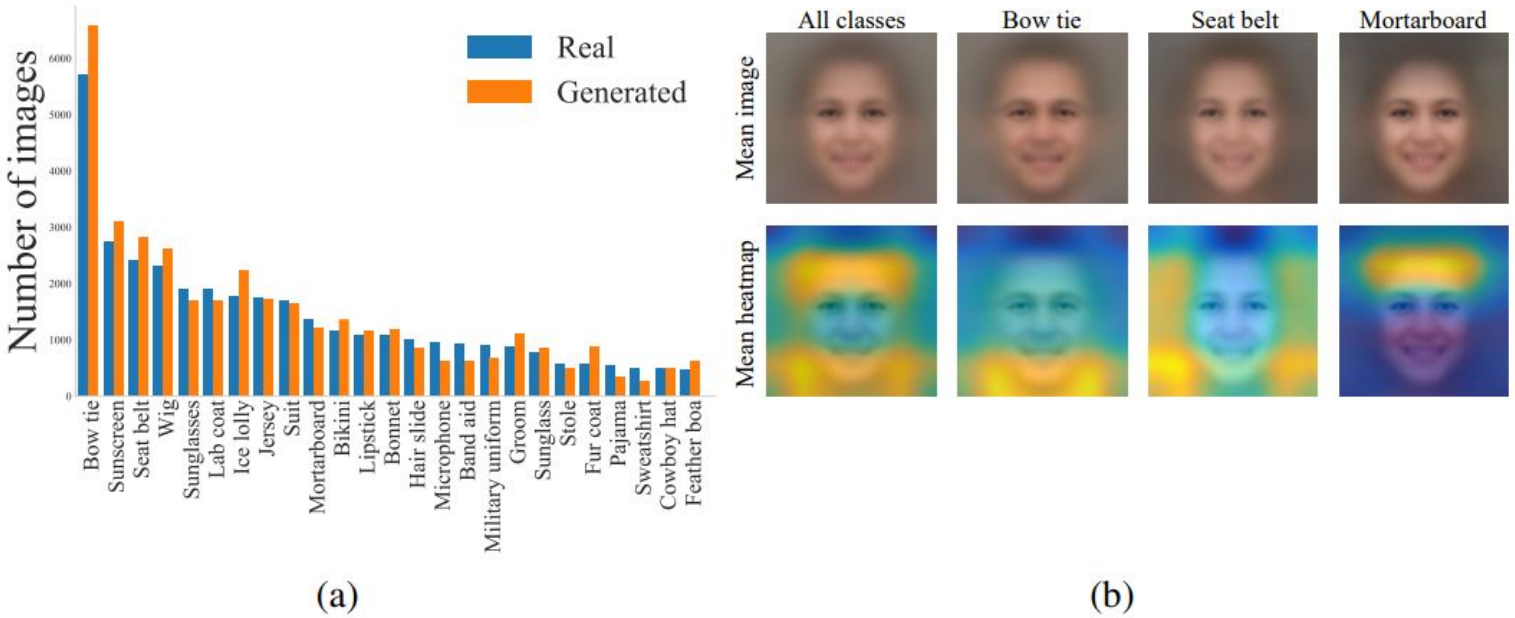


Figure 4: (a) Distribution of the ImageNet Top-1 classes, predicted by Inception-V3, for real and StyleGAN2 generated images in FFHQ. (b) Mean images and Grad-CAM heatmaps among all classes, and images classified as “bow tie”, “seat belt” and “mortarboard” categories. Strikingly, when averaging over all classes, FID is the most sensitive to ImageNet objects that are located outside the face area.

Why does it happen?

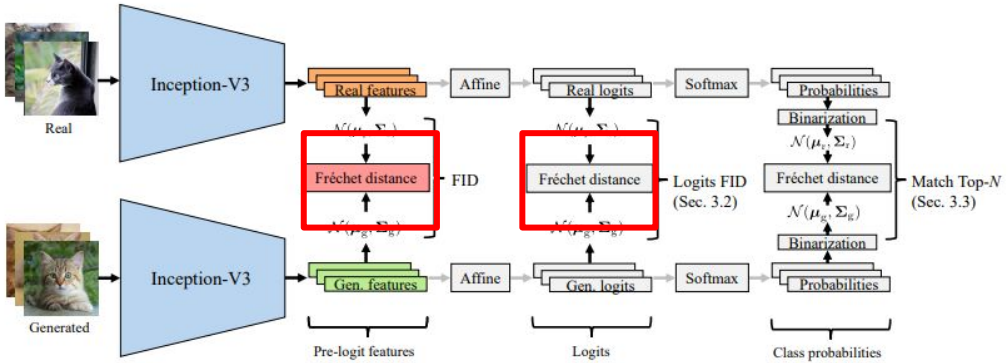


Figure 1: Overview of the Fréchet Inception Distance (FID) [16]. First, the real and generated images are separately passed through a pre-trained classifier network, typically the Inception-V3 [54], to produce two sets of feature vectors. Then, both distributions of features are approximated with multivariate Gaussians, and FID is defined as the Fréchet distance between the two Gaussians. In Section 3, we will compute alternative FIDs in the feature spaces of logits and class probabilities, instead of the usual pre-logit space.

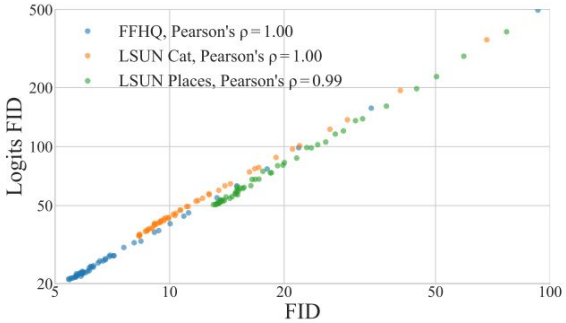


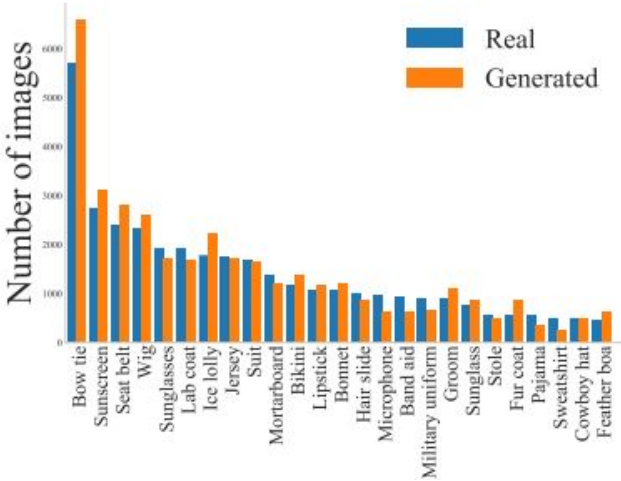
Figure 9: We observe nearly perfect correlation between FIDs computed from pre-logit features and classification logits since these two are separated by only one affine transformation. Each point corresponds to a single StyleGAN2 training snapshot in 256×256 resolution.

FID is essentially a distance between sets of ImageNet class probabilities.

Attacking FID: Top-1 Histogram Matching

Generated

Generated

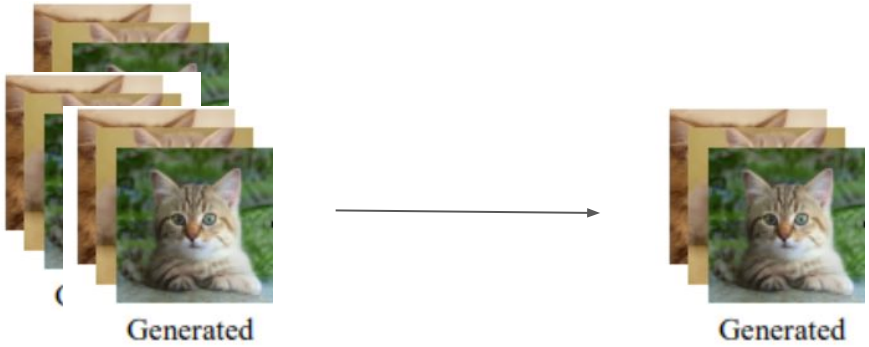


over-sample the fake images

remove the samples to match the Top-1 histogram with real dataset

Dataset	FID	FID^{Top-1}		FID_{CLIP}	FID^{Top-1}_{CLIP}	
FFHQ	5.63	4.91	(−12.8%)	2.90	2.86	(−1.4%)
LSUN CAT	8.35	7.37	(−11.7%)	8.95	8.83	(−1.3%)
LSUN CAR	5.75	5.17	(−10.1%)	7.76	7.73	(−0.4%)
LSUN PLACES	13.05	11.76	(−9.9%)	16.36	16.17	(−1.2%)
AFHQ-V2 DOG	10.58	9.39	(−11.2%)	4.23	4.16	(−1.7%)

Attacking FID: Weight Optimization



1. Over-sample the fake images (candidate images).
2. Assign non-negative weights to each sample.
3. Optimize the weights to minimize FID.
4. Use the weights as sampling probabilities and draw 50k random samples from candidate images.

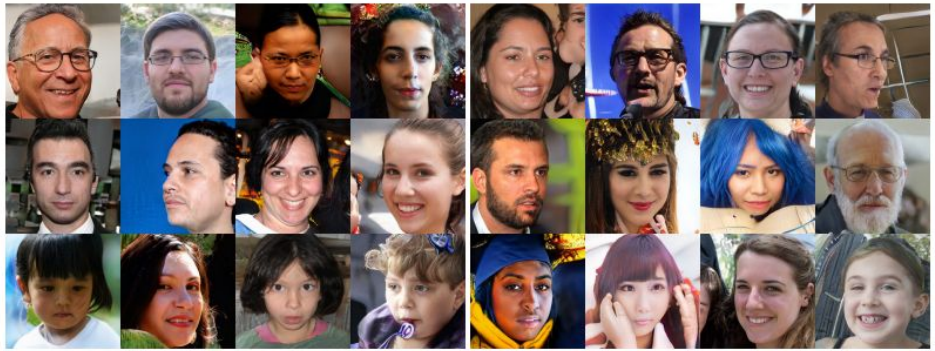
$$\min_{\boldsymbol{w}} \left(\left\| \boldsymbol{\mu}_{\mathrm{r}} - \boldsymbol{\mu}_{\mathrm{g}}(\boldsymbol{w}) \right\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_{\mathrm{r}} + \boldsymbol{\Sigma}_{\mathrm{g}}(\boldsymbol{w}) - 2\left(\boldsymbol{\Sigma}_{\mathrm{r}}\boldsymbol{\Sigma}_{\mathrm{g}}(\boldsymbol{w})\right)^{\frac{1}{2}} \right) \right), \tag{3}$$

$$\boldsymbol{\mu}_{\mathrm{g}}(\boldsymbol{w}) = \frac{\sum_i w_i \boldsymbol{f}_i}{\sum_i w_i} \text{ and } \boldsymbol{\Sigma}_{\mathrm{g}}(\boldsymbol{w}) = \frac{1}{\sum_i w_i} \sum_i w_i \left(\boldsymbol{f}_i - \boldsymbol{\mu}_{\mathrm{g}}(\boldsymbol{w})\right)^T \left(\boldsymbol{f}_i - \boldsymbol{\mu}_{\mathrm{g}}(\boldsymbol{w})\right)$$

Attacking FID: Weight Optimization

Table 2: Results of matching all fringe features. We compare the FID of randomly sampled images (FID) against ones that have been resampled to approximately match all fringe features with the training data (FID^{PL}). The numbers represent averages over ten FID evaluations. Additionally, we report the corresponding numbers by replacing the Inception-V3 feature space with CLIP features (FID_{CLIP}, FID^{PL}_{CLIP}). Note that we use CLIP only when computing FID; the resampling is still done using Inception-V3. The numerical values of FID and FID_{CLIP} are not comparable. The gray column (FID^{L}) shows an additional experiment where the resampling is done using logits instead of pre-logits.

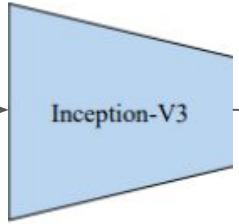
Dataset	FID	FID^{PL}		FID^{L}		FID_{CLIP}	FID^{PL}_{CLIP}	
FFHQ	5.63	1.82	(−67.7%)	2.31	(−59.0%)	2.90	2.78	(−4.1%)
LSUN CAT	8.35	3.05	(−63.5%)	3.88	(−53.5%)	8.95	7.98	(−10.8%)
LSUN CAR	5.75	2.11	(−63.3%)	2.59	(−55.0%)	7.76	7.33	(−5.5%)
LSUN PLACES	13.05	3.59	(−72.5%)	4.43	(−66.1%)	16.36	14.54	(−11.1%)
AFHQ-V2 DOG	10.58	5.92	(−44.0%)	6.27	(−40.7%)	4.23	4.04	(−4.5%)



(a) Images with small weights (b) Images with large weights

Figure 5: Uncurated random StyleGAN2 samples from images with (a) the smallest 10% of weights and (b) the largest 10% of weights after optimizing the weights to improve FID in the pre-logits space (Section 3.2). Both sets contain both realistic images and images with clear visual artifacts in apparently equal proportions, indicating that the large improvement in FID cannot be attributed to resampling simply assigning a lower weight to unrealistic images. The reader is encouraged to zoom in electronically. See supplementary material for a larger sample.

Attacking FID: Top-N histogram matching



[0.1, 0.2, 0.4, 0.3] -> [0, 0, 1, 1]
(Top-2 binarization)

minimize FID
(weight optimization)

[0, 0, 1, 1]
[0, 1, 1, 0]
[1, 0, 1, 0]
.
.
.

[0, 0, 1, 1]
[1, 0, 1, 0]
[0, 1, 1, 0]
.
.
.

Real

Generated

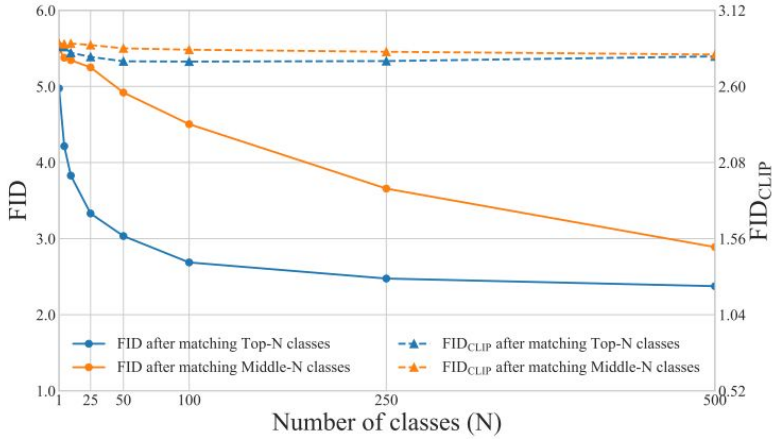
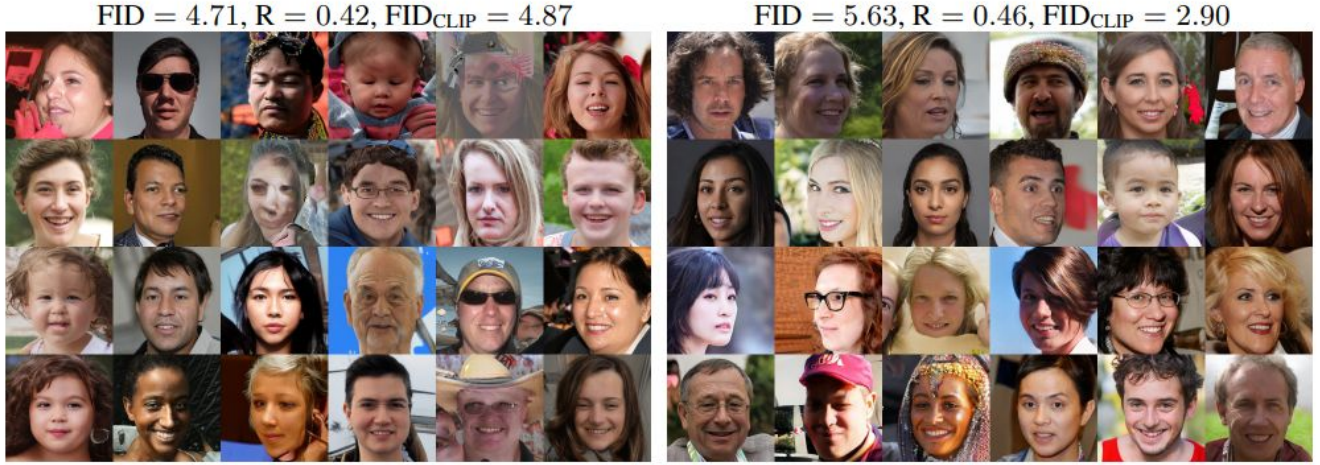


Figure 6: Softly matching Top-N class distributions in FFHQ through resampling (Section 3.3). FID (solid curves, left-hand y scale) decreases sharply with increasing number N of classes included in the Top-N indicator vectors, with $N = 10$ already yielding a highly significant improvement over the unoptimized model. At the same time, FID_{CLIP} (dashed curves, right-hand y scale) remains almost constant, indicating that the apparent improvements in FID are superfluous. The orange control curves have been computed from classes in the middle of the sorted probability vectors, indicating that the top classes indeed have a much stronger influence. As the numerical values of FID and FID_{CLIP} are not comparable, the left and right y axes have been normalized so that the unoptimized results are at the same height, and zero — would it be visible on the axes — would be on the same height. This ensures relative changes are represented accurately.



(a) Projected FastGAN (b) StyleGAN2

Figure 7: Uncurated samples from (a) Projected FastGAN and (b) StyleGAN2. While Projected FastGAN achieves a better FID, the samples contain clearly more artifacts. In contrast, Projected FastGAN has significantly higher FID$_{CLIP}$, consistent with the observed quality differential. This implies that the pre-trained features employed by Projected FastGAN have unintentional interference with (Inception-V3) FID.

Some models receive an unrealistically low FID simply by matching the ImageNet class distribution with training data. For instance, discriminator with ImageNet pre-trained feature extractor.

Conclusion

1. FID measures a distance between sets of ImageNet class probabilities.
2. It is possible to minimize FID without improving the quality simply by matching the class distribution.
3. ImageNet pre-training in generative model compromises the validity of FID. Be careful!
4. Maybe we should find alternative feature space (e.g., CLIP).