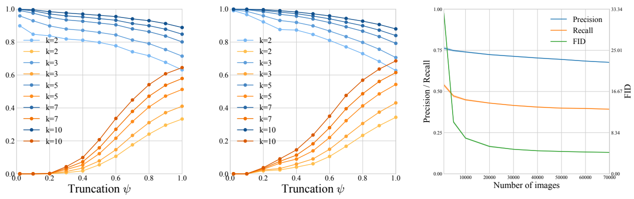
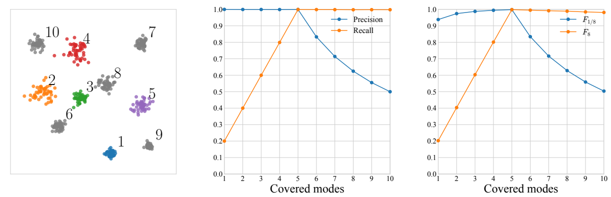


(a) Varying k, VGG-16 (b) Varying k, Inception-v3 (c) Varying $|\mathbf{\Phi}|$, VGG-16

Figure 1: (a) Precision (blue) and recall (orange) for several neighborhood sizes k. (b) Using Inception-v3 features instead of VGG-16 yields a substantially similar result. (c) Our metric behaves similarly to FID in terms of varying sample count.



(a) Data (b) Our method (c) Sajjadi et al. [1]

Figure 2: (a) Real data covers five modes (1–5) and the generated data is expanded, one mode at a time, to cover the real modes (1–5) and five extraneous modes (6–10). Both metrics were evaluated using 20k real and generated samples. (b) Results from our metric with $k = 3$. (c) Results from the method of Sajjadi et al. [1].

We thank the reviewers for their comments and remarks, and will gladly implement the suggested clarifications.

Reviewers 1 and 3 ask about different neighborhood sizes k, the number of samples $|\mathbf{\Phi}|$, and the choice of feature space. Figure 1a illustrates the effect of varying k in the setup used in Figure 4b of the submission (truncation sweep in StyleGAN, VGG-16 features, 50k samples). In general, different k yield consistent results and affect mainly the saturation towards 0 or 1. Therefore, selecting k is a tradeoff between under- or overestimating the manifolds. We chose $k = 3$ for slight underestimation, as overestimation leads to quicker saturation of precision and consequently makes it harder to measure differences between models. Figure 1b further shows that our metric is not sensitive to the choice of feature space: extracting the features from *pool3* of Inception-v3 [3] instead of VGG-16 makes no qualitative difference. Finally, Figure 1c shows that our metric behaves similarly to FID as the number of samples increases.

Reviewer 3 points out that our precision and recall do not measure the distance between generated and real distributions, and gives a counterexample where two continuous probability distributions have the same support sets but different densities. In this case our proposed metric would return perfect precision and recall scores, as it explicitly aims to disregard the density of the target distribution, measuring only the probability that a sample drawn from one distribution falls within the support of the other. FID remains an important tool for measuring distances between the distributions, and we argue that precision, recall, and FID all have well-justified roles in evaluating generative models as they provide complementary information about them.

Reviewer 3 further questions our claim that the curve representation in [1] is ambiguous. Making this claim was a result of an unfortunate grammatical mistake in our paper on line 57. Our intent was to say that the choice in [1] to use curves resulted *from* an ambiguity (that they discuss in Section 3.1), not that it came *with* any ambiguity. We apologize for the error and will revise the text. We have no objections to summarizing the curves using F_β scores as done in [1]. Furthermore, we thank the reviewer for bringing [2] to our attention, and will cite it as parallel work.

Finally, reviewer 3 suggested experimenting with simple or synthetic datasets similar to [1]. In Figure 2, we replicate the mode dropping and invention experiment in [1], albeit with a 10-class 2D Gaussian mixture model instead of CIFAR-10 images. As in [1], the real data covers five modes, and we measure precision and recall when 1–10 of the modes are covered by a hypothetical generator that draws samples from the corresponding Gaussian distributions. In Figure 2b we see that our method yields the correct values for precision and recall in all cases: when not all modes are being generated, precision is perfect and recall measures the fraction of modes covered, and when extraneous modes are generated, recall remains perfect while precision measures the fraction of real vs. generated modes. Figure 2c illustrates that the method of Sajjadi et al. [1] performs similarly except for artifacts from k-means clustering. We agree that including an experiment like this would strengthen the paper.

References

[1] M. S. M. Sajjadi, O. Bachem, M. Lucic, O. Bousquet, and S. Gelly. Assessing generative models via precision and recall. In *Proc. NIPS*, 2018.

[2] L. Simon, R. Webster, and J. Rabin. Revisiting precision and recall definition for generative model evaluation. *CoRR*, abs/1905.05441, 2019.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. CVPR*, 2015.