Supplementary Material: Evaluating generative networks using Gaussian mixtures of image features

Lorenzo Luzi[1,2]*, Carlos Ortiz Marrero[2], Nile Wynar[2], Richard G. Baraniuk[1], Michael J. Henry[2]
[1]Rice University, [2]Pacific Northwest National Laboratory†
{carlos.ortizmarrero,nile.wynar,michael.j.henry}@pnnl.gov
{enzo,richb}@rice.edu

A. Human evaluations with the BAPPS dataset

This experiment shows that both FID and WaM reflect human perception. To do this, we use the BAPPS [8] dataset which has human annotations. Specifically, we use the Two Alternative Forced Choice (2AFC) training set with traditional and convolutional neural network distortions. In BAPPS, a reference image is distorted in two different ways and the human evaluator must pick which of two distortions is closest to the reference image. There are two evaluators for each image, so the scores are either $0, 0.5$, or 1.0 indicating whether both evaluators picked the first image, each evaluator picked a different image, or both evaluators picked the second image, respectively.

We construct three different sub-datasets which capture varying qualities of images. Our "Good" dataset is composed of the images which both evaluators agree are better. Our "Bad" dataset is composed of the complementary images which both evaluators didn't pick. Our "Ambiguous" dataset is composed of the images for which the evaluators disagreed. We calculate FID and WaM by comparing the above datasets to the reference (noiseless) dataset. Hence, we would expect FID and WaM to have low values for the Good dataset, medium values for the Ambiguous dataset, and high values for the Bad dataset. Since these datasets have varying number of samples, we take subsets of the data to make all the datasets have $23,792$ samples to avoid biasing FID [2] differently for each dataset. Table 1 shows that both FID and WaM track with human perception.

Dataset	FID	WaM
Bad	28.8	90.3
Ambiguous	13.0	60.9
Good	7.0	50.4

Table 1: Experiment showing that both FID and WaM track with human perception using the BAPPS dataset. Both exhibit the ideal behavior: a decrease from Bad to Ambiguous to Good.

B. Human evaluations with the PIPAL dataset

This experiment shows that both FID and WaM reflect human perception. We use the PIPAL dataset [4] which has human annotations with finer distinctions between images than the BAPPS dataset [8] but with fewer samples. In constructing the PIPAL dataset, they employ the Elo rating system [3] to rank images on quality via human annotations. The training set has $23,200$ distorted images (from a 200 image reference set) which can be ranked from worst to best in quality based on their Elo scores.

We sort the dataset based on quality and then take subsets whose overall quality has a specific ordering. For example, if we use $n_{\text{bins}} = 10$, we make 10 sub-datasets with the first having the $2,320$ images with the lowest Elo score, the second having the $2,320$ images with the next lowest Elo scores, and so on. The only problem with this approach is that the number of images we have to work with is extremely limited, however this is the most appropriate dataset that exists to our knowledge. We use ResNet-18 to get features in $\mathbb{R}^{512}$ which allows the fitting of the Gaussian and GMM a little better; however, note that the reference dataset has only 200 samples, making this an extremely overparameterized problem. We calculate FID and WaM by comparing each sub-dataset to the reference (noiseless) dataset. Thus we average over 250 WaM values, modifying only the GMM initializations. FID does not need to be averaged over because it is deterministic. Our results are displayed in Figure 1.

C. GMMs can fit the distributions in Figure 1

Figure 1 of the main paper shows that Gaussians cannot model several distributions well. However, we show in Figure 2 that GMMs can indeed model those distributions.

*Work done while interning at Pacific Northwest National Laboratory
†Information Release Number: PNNL-SA-175469

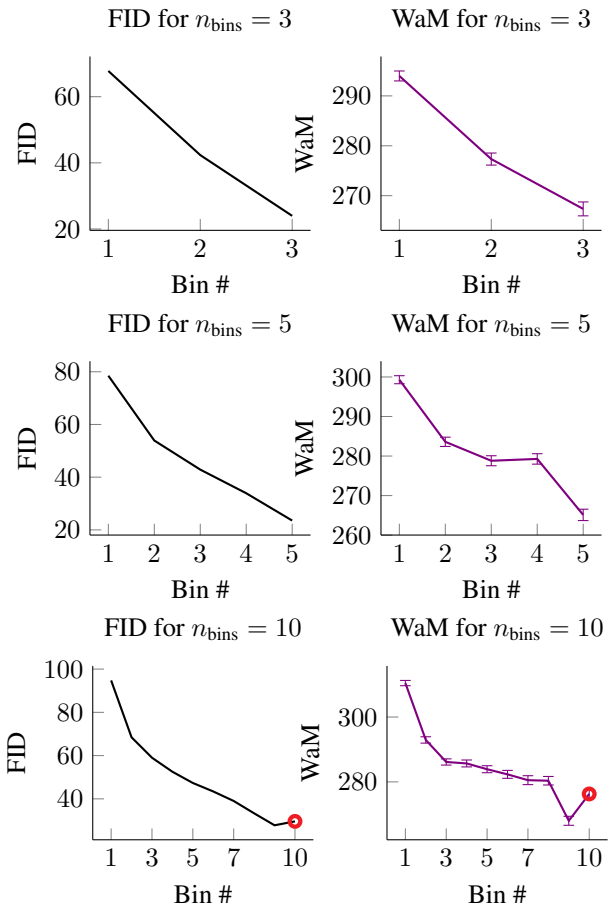

Figure 1: Experiment showing that both FID and WaM track with human perception using the PIPAL dataset. Both metrics should decrease monotonically and hence start to break down when we use $n_{\text{bins}} = 10$. The red circles indicate a statistically significant break in monotonicity.

D. Targeted perturbations (extended)

Here we show the targeted perturbation results in more detail than in Section 5.1 of the main paper. We don't show figures of the images before and after perturbation besides Figure 4 of the main paper because they are all imperceptible, with maximum pixel differences of 0.25%. All the FID, WaM, and R values were calculated using Inception-v3. We use Equation (3) for Table 2, Equation (4) for Table 3, Equation (5) for Table 4, Equation (6) for Table 5, and FID for Table 6. We follow recent work [6, 5][1] in order to backpropagate through FID. Our results show that in every case, FID is significantly more sensitive to imperceptible perturbations of the first two moments when compared to WaM.

[1]Although the authors of the paper introduced a Fast FID, we backpropagate through FID in our work.

E. Kernel Inception distance experiments

Kernel Inception distance (KID) [1] is a popular method to evaluate the performance of a GAN which uses embeddings from powerful classifiers, such as Inception-v3 [7]. We use the cubic polynomial kernel, i.e., $k(\boldsymbol{x}, \boldsymbol{y}) = (\frac{1}{d}\boldsymbol{x}^\top \boldsymbol{y} + 1)^3$ for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, to compute similarities between featurized samples, as is typically done. We use this method to evaluate WaM's sensitivity to imperceptible noise perturbations. To do this, we define R_{KID} to be the ratio of the KID of the perturbed images over the KID of the original images. We further define

$$R' = \frac{R_{\text{KID}}}{R_{\text{WaM}}}.$$

All the KID, WaM, and R values were calculated using Inception-v3. We use Equation (3) for Table 7, Equation (4) for Table 8, Equation (5) for Table 9, Equation (6)for Table 10, and FID for Table 11. These results show that KID is is still significantly affected by these perturbations, even though some values of R_{KID} are smaller than R_{WaM}. WaM is less sensitive than both FID and KID in the majority of these experiments, implying that it does not depend as heavily on the first two moments and can capture more higher order information than both metrics.

We now consider the random perturbations in Section 5.2 in the main paper and evaluate R' on them, as shown in Tables 12 and 13. We see that KID has similar sensitivity to WaM on BigGAN generated images but much higher sensitivity on real images. In fact, KID has higher sensitivity on real images than FID. We stress that the ability to evaluate

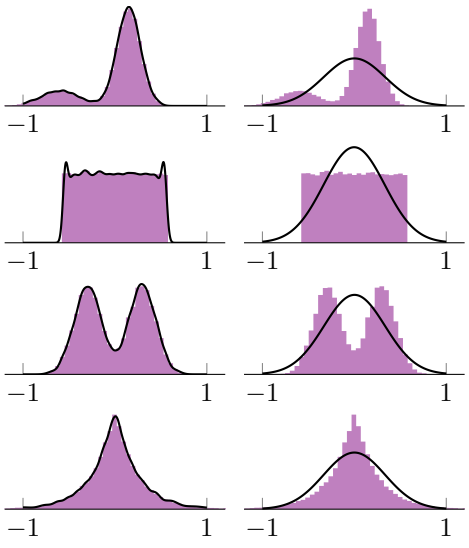

Figure 2: GMMs are able to fit a variety of distributions significantly better than Gaussians. We use a GMM with $k = 20$ components.

realistic images is important because that is what we **want** to generate. Therefore, WaM provides a means to evaluate realistic images better than FID and KID under imperceptible noise perturbations.

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
$\text{FID} = 55.7$	$\text{FID} = 154.2$	$\text{FID} = 3.7$	$\text{FID} = 46.6$
$\text{WaM}^2 = 378.4$	$\text{WaM}^2 = 424.3$	$\text{WaM}^2 = 237.0$	$\text{WaM}^2 = 280.0$
$R_{\text{FID}} = 2.8$		$R_{\text{FID}} = 12.7$	
$R_{\text{WaM}} = 1.1$		$R_{\text{WaM}} = 1.2$	
$R = \mathbf{2.5}$		$R = \mathbf{10.8}$	

Table 2: Mean perturbations: We show that FID values are significantly more sensitive to imperceptible perturbations to the feature means (Equation (3)).

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
$\text{FID} = 55.7$	$\text{FID} = 125.7$	$\text{FID} = 3.7$	$\text{FID} = 113.0$
$\text{WaM}^2 = 378.4$	$\text{WaM}^2 = 540.8$	$\text{WaM}^2 = 237.0$	$\text{WaM}^2 = 422.5$
$R_{\text{FID}} = 2.3$		$R_{\text{FID}} = 30.9$	
$R_{\text{WaM}} = 1.4$		$R_{\text{WaM}} = 1.8$	
$R = \mathbf{1.6}$		$R = \mathbf{17.3}$	

Table 3: Covariance perturbations: We show that FID values are significantly more sensitive to imperceptible perturbations to the feature covariances (Equation (4)).

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
$\text{FID} = 55.7$	$\text{FID} = 177.6$	$\text{FID} = 3.7$	$\text{FID} = 106.9$
$\text{WaM}^2 = 378.4$	$\text{WaM}^2 = 521.3$	$\text{WaM}^2 = 237.0$	$\text{WaM}^2 = 412.2$
$R_{\text{FID}} = 3.2$		$R_{\text{FID}} = 29.2$	
$R_{\text{WaM}} = 1.4$		$R_{\text{WaM}} = 1.7$	
$R = \mathbf{2.3}$		$R = \mathbf{16.8}$	

Table 4: Mean-covariance perturbations: We show that FID values are significantly more sensitive to imperceptible perturbations to the feature means and covariances together (Equation (5)).

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
$\text{FID} = 55.7$	$\text{FID} = 145.9$	$\text{FID} = 3.7$	$\text{FID} = 112.2$
$\text{WaM}^2 = 378.4$	$\text{WaM}^2 = 578.3$	$\text{WaM}^2 = 237.0$	$\text{WaM}^2 = 444.0$
$R_{\text{FID}} = 2.6$		$R_{\text{FID}} = 30.7$	
$R_{\text{WaM}} = 1.5$		$R_{\text{WaM}} = 1.9$	
$R = \mathbf{1.7}$		$R = \mathbf{16.4}$	

Table 5: Alternative covariance perturbations: We show that FID values are significantly more sensitive to imperceptible perturbations to the feature covariances, using a different metric on the covariances than the Frobenius norm (Equation (6)).

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
FID $= 55.7$	FID $= 166.5$	FID $= 3.7$	FID $= 112.0$
$\text{WaM}^2 = 378.4$	$\text{WaM}^2 = 548.5$	$\text{WaM}^2 = 237.0$	$\text{WaM}^2 = 377.0$
$R_{\text{FID}} = 3.0$		$R_{\text{FID}} = 30.6$	
$R_{\text{WaM}} = 1.4$		$R_{\text{WaM}} = 1.6$	
$R = \mathbf{2.1}$		$R = \mathbf{19.2}$	

Table 6: FID perturbations: We show that FID values are significantly more sensitive to imperceptible perturbations when we adversarially attempt to inflate FID.

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
KID $= 0.029$	KID $= 0.139$	KID $= 0.0007$	KID $= 0.066$
$\text{WaM}^2 = 378.4$	$\text{WaM}^2 = 424.3$	$\text{WaM}^2 = 237.0$	$\text{WaM}^2 = 280.0$
$R_{\text{KID}} = 4.7$		$R_{\text{KID}} = 94.6$	
$R_{\text{WaM}} = 1.1$		$R_{\text{WaM}} = 1.2$	
$R' = \mathbf{4.2}$		$R' = \mathbf{80.1}$	

Table 7: Mean perturbations: We show that KID values are significantly more sensitive to imperceptible perturbations to the feature means (Equation (3)).

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
KID $= 0.029$	KID $= 0.014$	KID $= 0.0007$	KID $= 0.087$
$\text{WaM}^2 = 378.4$	$\text{WaM}^2 = 540.8$	$\text{WaM}^2 = 237.0$	$\text{WaM}^2 = 422.5$
$R_{\text{KID}} = 0.5$		$R_{\text{KID}} = 125.6$	
$R_{\text{WaM}} = 1.4$		$R_{\text{WaM}} = 1.8$	
$R' = \mathbf{0.3}$		$R' = \mathbf{70.5}$	

Table 8: Covariance perturbations: We show that KID values are significantly more sensitive to imperceptible perturbations to the feature covariances (Equation (4)).

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
KID $= 0.029$	KID $= 0.097$	KID $= 0.0007$	KID $= 0.100$
$\text{WaM}^2 = 378.4$	$\text{WaM}^2 = 521.3$	$\text{WaM}^2 = 237.0$	$\text{WaM}^2 = 412.2$
$R_{\text{KID}} = 3.3$		$R_{\text{KID}} = 143.9$	
$R_{\text{WaM}} = 1.4$		$R_{\text{WaM}} = 1.7$	
$R' = \mathbf{2.4}$		$R' = \mathbf{80.8}$	

Table 9: Mean-covariance perturbations: We show that KID values are significantly more sensitive to imperceptible perturbations to the feature means and covariances together (Equation (5)).

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
KID $= 0.029$	KID $= 0.034$	KID $= 0.0007$	KID $= 0.074$
WaM$^2 = 378.4$	WaM$^2 = 578.3$	WaM$^2 = 237.0$	WaM$^2 = 444.0$
$R_{\text{KID}} = 1.2$		$R_{\text{KID}} = 106.1$	
$R_{\text{WaM}} = 1.5$		$R_{\text{WaM}} = 1.9$	
$R' = \mathbf{0.8}$		$R' = \mathbf{56.6}$	

Table 10: Alternative covariance perturbations: We show that KID values are significantly more sensitive to imperceptible perturbations to the feature covariances, using a different metric on the covariances than the Frobenius norm (Equation (6)).

Original (BigGAN)	Perturbed (BigGAN)	Original (ImageNet)	Perturbed (ImageNet)
KID $= 0.029$	KID $= 0.057$	KID $= 0.0007$	KID $= 0.077$
WaM$^2 = 378.4$	WaM$^2 = 548.5$	WaM$^2 = 237.0$	WaM$^2 = 377.0$
$R_{\text{KID}} = 2.0$		$R_{\text{KID}} = 111.5$	
$R_{\text{WaM}} = 1.4$		$R_{\text{WaM}} = 1.6$	
$R' = \mathbf{1.3}$		$R' = \mathbf{70.1}$	

Table 11: FID perturbations: We show KID values are significantly more sensitive to imperceptible perturbations when we adversarially attempt to inflate FID.

	$\sigma = 0.01$	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.5$
KID(orig)	2.11	2.11	2.11	2.11	2.11
KID(pert)	2.22	2.74	3.47	4.65	5.23
WaM^2(orig)	504.30	504.30	504.30	504.30	504.30
WaM^2(pert)	539.54	516.75	628.68	748.65	1328.01
R_{KID}	1.05	1.29	1.64	2.2	2.47
R_{WaM}	1.07	1.02	1.25	1.48	2.63
R'	**0.98**	**1.26**	**1.31**	**1.49**	**0.94**

Table 12: R' values for BigGAN-generated images using additive isotropic Gaussian noise (as explained in Section 5.2 of the main paper)showing that KID has similar sensitivity as WaM to noise perturbations of generated images. The original image above was randomly selected from a set of $50{,}000$ images generated by BigGAN. The KID, WaM, and R' values were calculated using ResNet-18.

	$\sigma = 0.01$	$\sigma = 0.05$	$\sigma = 0.1$	$\sigma = 0.2$	$\sigma = 0.5$
KID(orig)	0.025	0.025	0.025	0.025	0.025
KID(pert)	0.033	0.187	0.496	1.146	2.745
WaM^2(orig)	208.45	208.45	208.45	208.45	208.45
WaM^2(pert)	219.49	316.06	549.03	1081.28	4007.29
R_{KID}	1.283	7.363	19.528	45.143	108.118
R_{WaM}	1.05	1.52	2.63	5.19	19.22
R'	**1.22**	**4.84**	**7.43**	**8.70**	**5.63**

Table 13: R' values for real images (ImageNet validation data) using additive isotropic Gaussian noise (as explained in Section 5.2 of the main paper) showing that KID is more sensitive than WaM to noise perturbations of real images. The original image above was randomly selected from a set of $50{,}000$ images of the ImageNet validation dataset. In contrast to Figure 5 of the main paper, we see that KID is more sensitive to these perturbations when the images look more realistic. The FID and WaM values were calculated using ResNet-18.

References

[1] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. *arXiv preprint arXiv:1801.01401*, 2018.

[2] Min Jin Chong and David Forsyth. Effectively unbiased fid and inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6070–6079, 2020.

[3] Arpad E Elo. *The rating of chessplayers, past and present*. Arco Pub., 1978.

[4] Gu Jinjin, Cai Haoming, Chen Haoyu, Ye Xiaoxing, Jimmy S Ren, and Dong Chao. Pipal: a large-scale image quality assessment dataset for perceptual image restoration. In *European Conference on Computer Vision (ECCV) 2020*, pages 633–651, Cham, 2020. Springer International Publishing.

[5] Alexander Mathiasen and Frederik Hvilshøj. Backpropagating through fr\'echet inception distance. *arXiv preprint arXiv:2009.14075*, 2020.

[6] Alexander Mathiasen and Frederik Hvilshøj. Fast fr\'echet inception distance. *arXiv preprint arXiv:2009.14075*, 2020.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[8] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.