Quantitatively Evaluating GANs with Divergence Proposed for Training

Daniel Jiwoong Im[1,2], He Ma[3,4], Graham W. Taylor[3,4], Kristin Branson[1]   [1]Janelia Research Campus, [2]AIFounded, [3]University of Guelph, [4]Vector Institute

Motivations and Contributions

- It is difficult to compare GAN models and understand their relative strengths and weaknesses because we lack quantitative methods for assessing the learned generators.
- We evaluate the performance of various types of GANs using divergence and distance functions typically used for training.
- We qualitatively and quantitatively compared these metrics to human perception, and found that our proposed metrics better reflected human perception.

Evaluation Metrics

We consider the following four metrics that are commonly used to train GANs:

Let $D_\phi : \mathcal{X} \rightarrow \{0, 1\}$ be the discriminator and $G_\theta : \mathcal{Z} \rightarrow \mathcal{X}$ be the generator.

Original GAN Criterion (GC)
Training a standard GAN corresponds to the following:

$$\max_\phi \quad \mathbb{E}_{x\sim p(x)}[\log(D_\phi(x))] + \mathbb{E}_{z\sim p(z)}[\log(1 - D_\phi(G_\theta(z)))].$$

Least-Squares GAN Criterion (LS)
A Least-Squares GAN corresponds to training with a Pearson χ^2 divergence:

$$\max_\phi \quad -\mathbb{E}_{x\sim p(x)}[(D_\phi(x) - b)^2] - \mathbb{E}_{z\sim p(z)}[(D_\phi(G_\theta(z) - a))^2].$$

We set $a = 0$ and $b = 1$ when training D_ϕ.

Maximum Mean Discrepancy (MMD)
MMD considers the largest difference in the expectations over a unit ball of RKHS $\mathcal{H}$ with with kernel $k(\cdot, \cdot)$.

$$\begin{aligned}\max_{\phi:\|f\|_{\mathcal{H}}\leq 1} & \mathbb{E}_{x\sim p(x)}[D_\phi(x)] - \mathbb{E}_{z\sim p(z)}[D_\phi(G_\theta(z)] \\ &= \mathbb{E}_{x,x'\sim P}\left[k(x, x')\right] + \mathbb{E}_{z,z'\sim p(Z)}\left[k(G_\theta(z), G_\theta(z'))\right] \\ &\quad - 2\mathbb{E}_{x\sim P, z\sim p(Z)}\left[k(x, G_\theta(z))\right].\end{aligned}$$

Improved Wasserstein Distance (IW) The dual form of the Wasserstein distance for training GANs.

$$\max_{\phi:\|\phi\|_L\leq 1} \left[\mathbb{E}_{x\sim p(X)}\left[D_\phi(x)\right] - \mathbb{E}_{z\sim p(Z)}\left[D_\phi(G_\theta(z))\right]\right].$$

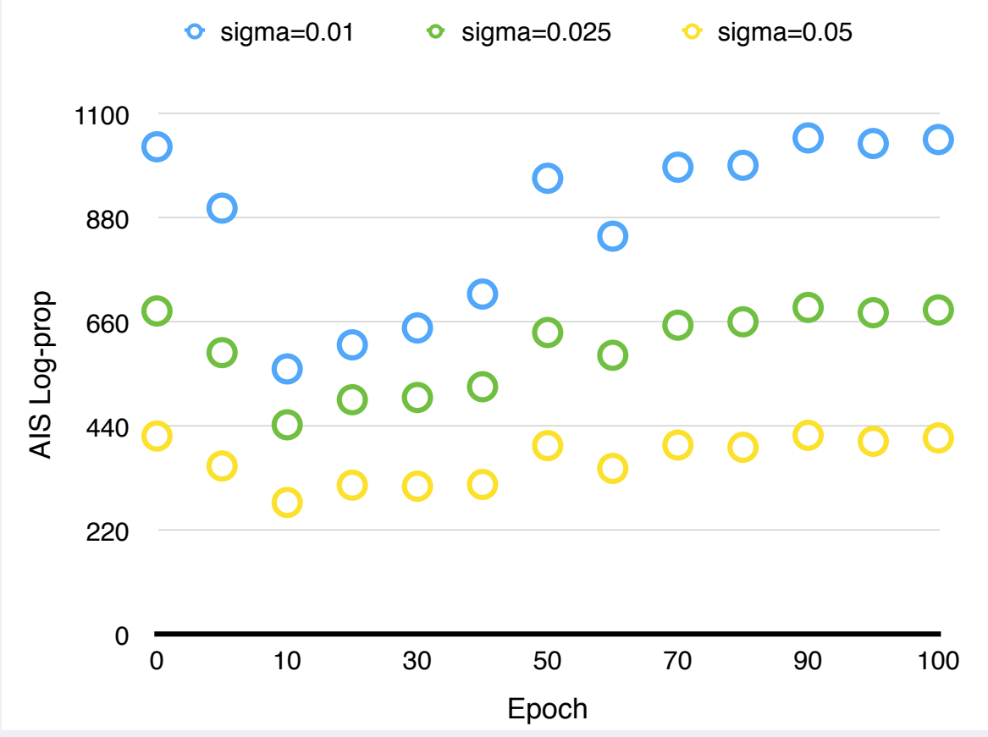
Test Procedure

Let G_θ be the generator to be evaluated, X_{tr} be the training data, and $J(\{x_m\}, \{s_m\}; \phi)$ be the divergence/distance.

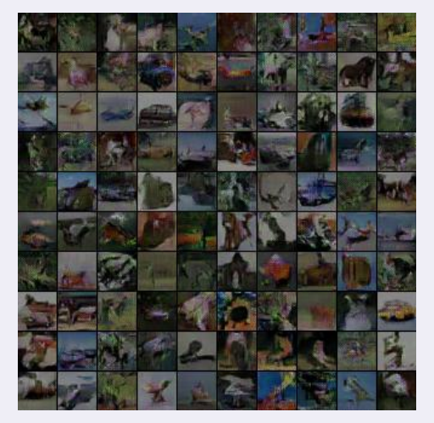
```
1. Initialize critic's parameter $\phi$
2. For $i = 1 \cdots N$
       Sample data points from X, $\{x_m\} \sim X_{tr}$.
       Sample points from generator, $\{s_m\} \sim G_\theta$.
       $\phi \leftarrow \phi + \eta \nabla_\phi J(\{x_m\}, \{s_m\}; \phi)$.
3. Sample points from generative model, $\{s_m\} \sim G_\theta$.
4. return $J(X_{te}, \{s_m\}; \phi)$.
```

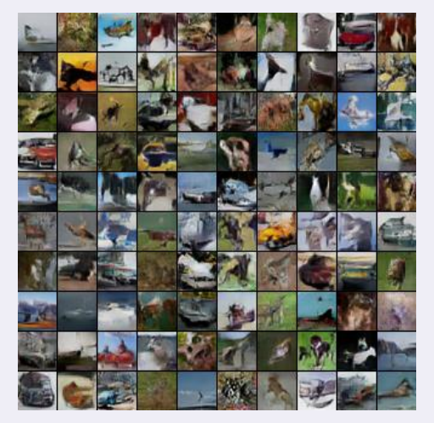
Observation on existing metrics

Log-likelihood estimated uinsg Annealed Importance Sampling (AIS) for generators learned using DCGAN at various points during training, MNIST data set.

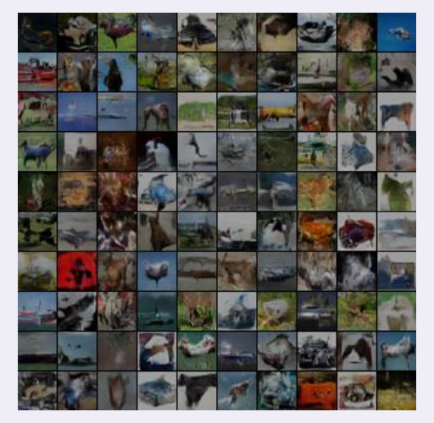


- We observe a high log-likelihood at the beginning of training, followed by a drop in likelihood, which then returns to the high value.
- MMD is overly sensitive to image intensity, while Inception Score (IS) is under-sensitive to image intensity.


IS = 6.45


IS = 6.31

IS= 5.11
MMD= 0.03


IS= 6.15
MMD= 0.49

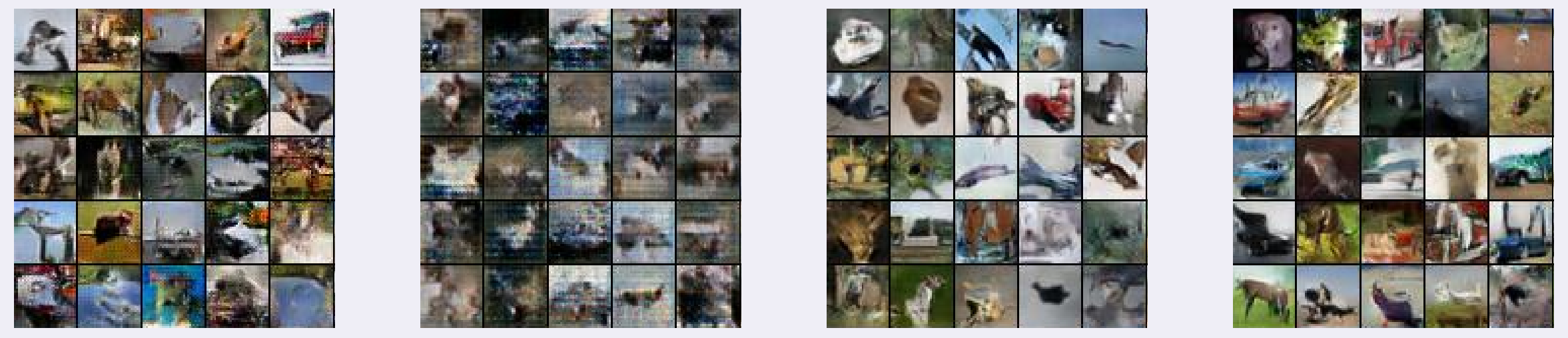
Hyperparamter Analysis

LS-DCGAN and W-DCGAN scores on CIFAR10 w.r.t different generator and discriminator size.

Model	Architecture Feature Map	MMD Test vs. Samples	IW
W-DCGAN	(e)	0.1057 ± 0.0798	450.17± 25.74
	(f)	0.2176 ± 0.2706	16.52 ± 15.63
LS-DCGAN	(e)	0.1390 ± 0.1525	343.23± 47.55
	(f)	0.0054 ± 0.0022	12.75 ± 4.29

Model	Architecture Feature Map	LS	IS (ResNet)
W-DCGAN	(e)	-0.0079 ± 0.0009	6.403 ± 0.839
	(f)	-0.0636 ± 0.0101	6.266 ± 0.055
LS-DCGAN	(e)	-0.0092 ± 0.0007	5.751 ± 0.511
	(f)	-0.0372 ± 0.0068	6.600 ± 0.061

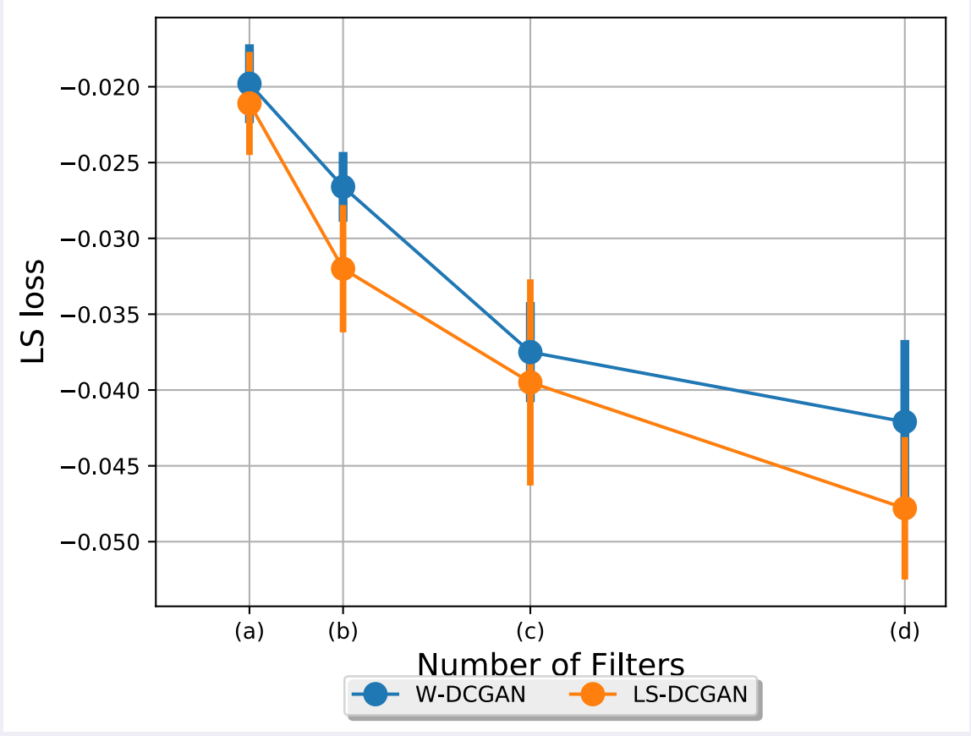
Samples from different LS-DCGAN architectures.


(a) (e) (c) (d)

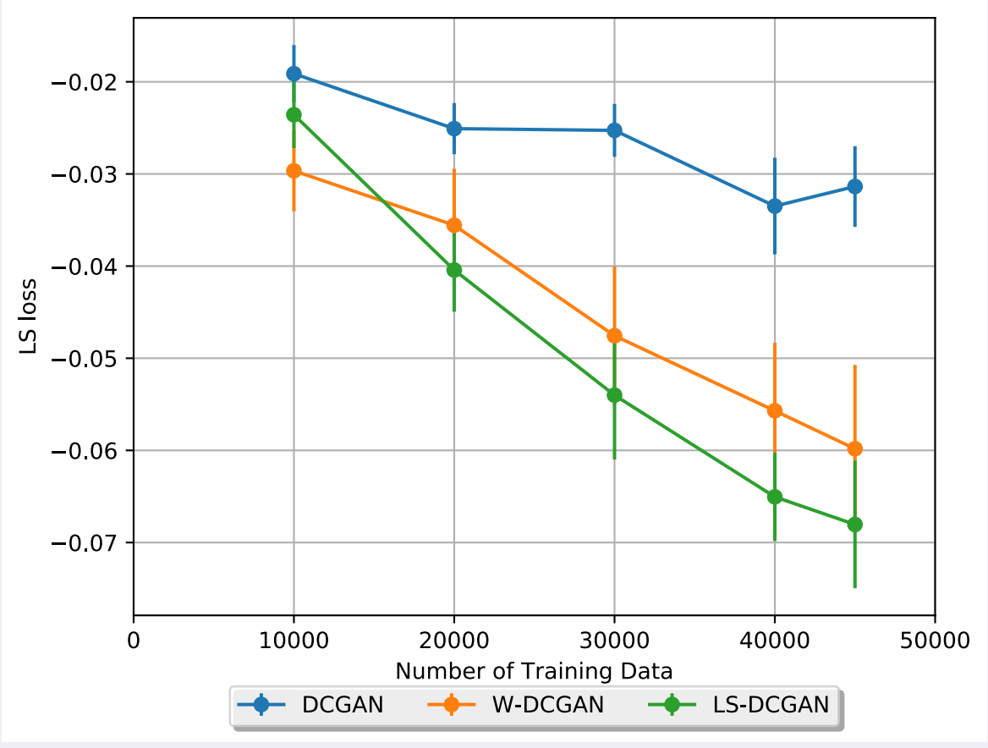
LS-DCGAN and W-DCGAN scores on CIFAR10 w.r.t the dimensionality of the noise vector.

$\mid z \mid$	LS-DCGAN		W-DCGAN	
	IW	LS	IW	LS
50	3.9010 ± 0.60	-0.0547 ± 0.0059	6.0948 ± 3.21	-0.0532 ± 0.0069
100	5.6588 ± 1.47	-0.0511 ± 0.0065	5.7358 ± 3.25	-0.0528 ± 0.0051
150	5.8350 ± 0.80	-0.0434 ±0.0036	3.6945 ± 1.33	-0.0521 ± 0.0050

LS score evaluation



w.r.t feature map size.



w.r.t amount of training data.

Metric Comparisons

GAN scores for various metrics trained on MNIST.

Model	MMD ↓	IW ↓	GC ↓	LS ↓	IS (Logistic Reg.) ↑
DCGAN	0.028 ± 0.0066	7.01 ± 1.63	-2.2e-3 ± 3e-4	-0.12 ± 0.013	5.76 ± 0.10
W-DCGAN	0.006 ± 0.0009	7.71 ± 1.89	-4e-4 ± 4e-4	-0.05 ± 0.008	5.17 ± 0.11
LS-DCGAN	0.012 ± 0.0036	4.50 ± 1.94	-3e-3 ± 6e-4	-0.13 ± 0.022	6.07 ± 0.08

Lighter color indicates better performance.

GAN scores for various metrics trained on CIFAR10.

Model	MMD	IW	LS	IS (ResNet)	FID
DCGAN	0.0538 ± 0.014	8.844 ± 2.87	-0.0408 ± 0.0039	6.649 ± 0.068	0.112 ± 0.010
W-DCGAN	0.0060 ± 0.001	9.875 ± 3.42	-0.0421 ± 0.0054	6.524 ± 0.078	0.095 ± 0.003
LS-DCGAN	0.0072 ± 0.0024	7.10 ± 2.05	-0.0535 ± 0.0031	6.761 ± 0.069	0.088 ± 0.008

Reference for the different architectures explored in the experiments.

Label	Feature Maps	
	Discriminator	Generator
(a)	[3, 16 , 32 , 64]	[128 , 64 , 32 , 3]
(b)	[3, 32 , 64 , 128]	[256 , 128, 64 , 3]
(c)	[3, 64 , 128, 256]	[512 , 256, 128, 3]
(d)	[3, 128, 256, 512]	[1024, 512, 256, 3]
(e)	[3, 16 , 32 , 64]	[1024, 512, 256, 3]
(f)	[3, 128, 256, 512]	[128 , 64 , 32 , 3]

GAN scores for various metrics trained on LSUN Bedroom dataset.

Model	MMD	IW	LS
DCGAN	0.00708	3.79097	-0.14614
W-DCGAN	0.00584	2.91787	-0.20572
LS-DCGAN	0.00973	3.36779	-0.17307

Evaluation of GANs on MNIST and Fashion-MNIST datasets.

Model	MNIST			Fashion-MNIST		
	IW	LS	FID	IW	LS	FID
DCGAN	0.4814 ± 0.0083	-0.111 ± 0.0074	1.84 ± 0.15	0.69 ± 0.0057	-0.0202 ± 0.00242	3.23 ± 0.34
EBGAN	0.7277 ± 0.0159	-0.029 ± 0.0026	5.36 ± 0.32	0.99 ± 0.0001	-2.2e-5 ± 5.3e-5	104.08 ± 0.56
W-DCGAN GP	0.7314 ± 0.0194	-0.035 ± 0.0059	2.67 ± 0.15	0.89 ± 0.0086	-0.0005 ± 0.00037	2.56 ± 0.25
LS-DCGAN	0.5058 ± 0.0117	-0.115 ± 0.0070	2.20 ± 0.27	0.68 ± 0.0086	-0.0208 ± 0.00290	0.62 ± 0.13
BEGAN	-	-0.009 ± 0.0063	15.9 ± 0.48	0.90 ± 0.0159	-0.0016 ± 0.00047	1.51 ± 0.16
DRAGAN	0.4632 ± 0.0247	-0.116 ± 0.0116	1.09 ± 0.13	0.66 ± 0.0108	-0.0219 ± 0.00232	0.97 ± 0.14

- Every metric, except for MMD, showed that LS-DCGAN performed best for MNIST and CIFAR10, while W-DCGAN performed best for LSUN.
- The standard deviations for the IW distance are higher than for LS divergence.
- We found that the different GAN frameworks have significantly different performance according to the LS-GAN criterion, but not according to the IW criterion (p < .05, Wilcoxon rank-sum test). Thus LS is more sensitive than IW.

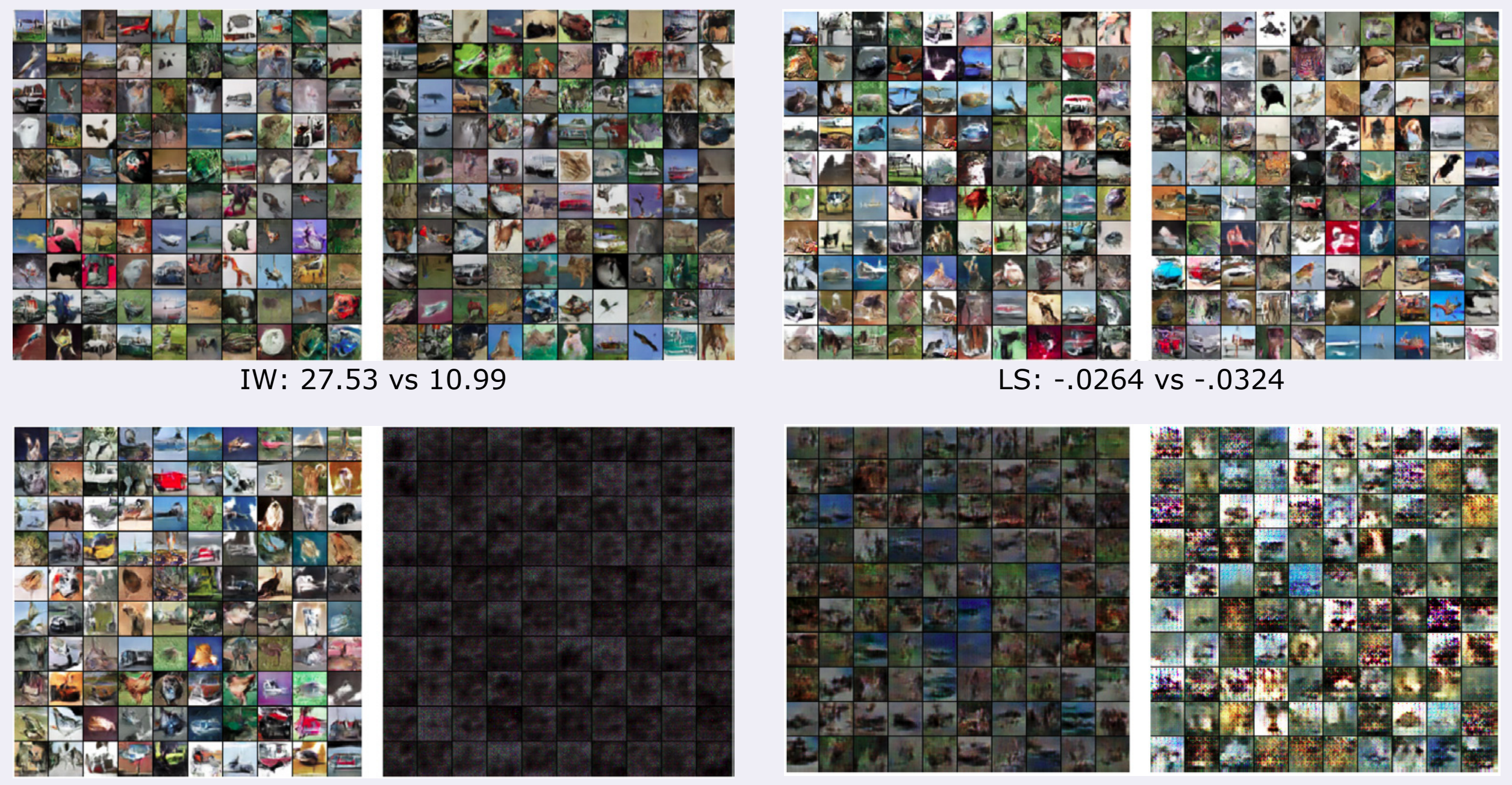
Comparison to Human Perception

We compared the LS, IW, MMD, and IS metrics to human perception for the CIFAR10 dataset.

We asked five volunteers to choose which of two sets of 100 samples, each generated using a different generator, looked most realistic.

Metric	Fraction	[Agreed/Total] samples	p<.05?
IW	0.977	128 / 131	* *
LS	0.931	122 / 131	*
IS	0.863	113 / 131	*
MMD	0.832	109 / 131	* *

- Table presents the fraction of pairs for which each metric agrees with humans (higher is better). The fraction of pairs of which each metric agrees with human scores. We use colored asterisks to represent significant differences (two-sided Fisher's test, $p < .05$). E.g. * in the IW row indicates that IW and IS are significantly different.
- IW has a slight edge over LS, and both outperform IS and MMD.


IW: 27.53 vs 10.99
LS: -.0264 vs -.0324
IS: 6.22 vs 7.89
MMD: 1.19 vs 0.13

Pairs of generated image sets for which human perception and metrics disagree. We selected one such example for each metric for which the difference in that metric's scores was high. For each pair, humans perceived the set of images on the left to be more realistic than those on the right, while the metric predicted the opposite.