REAL OR NOT REAL, THAT IS THE QUESTION

ICLR 2020

Yuanbo Xiangli1∗ , Yubin Deng1∗ , Bo Dai1∗ , Chen Change Loy2 , Dahua Lin1

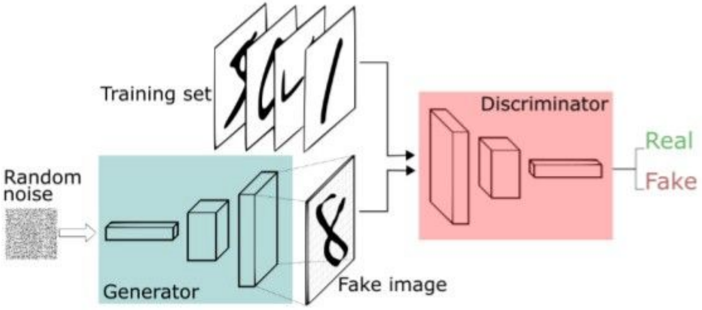
Abstract

- Generalize the standard GAN to a new perspective by treating realness as a random variable that can be estimated from multiple angles.
- In this generalized framework, referred to as RealnessGAN , the discriminator outputs a distribution as the measure of realness
- Compared to multiple baselines, RealnessGAN provides stronger guidance for the generator, achieving improvements on both synthetic and real-world datasets

Background - GAN

Generative Adversarial Networks (GANs)

GAN training procedure pits two neural networks against each other, a generator and a discriminator.



$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))].$$

log prob of D predicting that real-world data is genuine

log prob of D predicting that G's generated data is not genuine

$$\begin{aligned} L_D &= -\mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] - \mathbb{E}_{z\sim p_z}[1 - \log D(G(z))] \\ L_G &= -\mathbb{E}_{z\sim p_z}[\log D(G(z))] \end{aligned}$$

Introduction

- In the standard formulation (Goodfellow et al., 2014), the realness of an input sample is estimated by the discriminator using a single scalar.
- However, for high dimensional data such as images, we naturally perceive them from more than one angles and deduce whether it is life-like based on multiple criteria.
- In this paper they propose to **generalize the standard framework (Goodfellow et al., 2014) by treating realness as a random variable, represented as a distribution rather than a single scalar**

Figure 1: The perception of realness depends on various aspects. (a) Human-perceived flawless. (b) Potentially reduced realness due to: inharmonious facial structure/components, unnatural background, abnormal style combination and texture distortion.

A Distributional view of Realness

- substituting the scalar output of a discriminator D with a distribution $P_{realness}$

$$D(\boldsymbol{x}) = \{p_{\text{realness}}(\boldsymbol{x}, u); u \in \Omega\}$$

- Where Ω is the set of outcomes of $P_{realness}$ and each outcome u can be viewed as a potential realness measure
- While the standard GAN used two virtual ground-truth scalars, in our case two virtual ground-truth distributions are A_1 (real) A_0 (fake) which are also defined on Ω will represent real/fake.

$$\max_{G} \min_{D} V(G, D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(\boldsymbol{x}))].$$

A Distributional view of Realness

- <u>Note:</u> if we let $P_{realness}$ be a discrete distribution with two outcomes $\{u_0, u_1\}$, and set:

  $$\mathcal{A}_0(u_0) = \mathcal{A}_1(u_1) = 1 \qquad \mathcal{A}_0(u_1) = \mathcal{A}_1(u_0) = 0,$$

  the updated objective in equation 3 can be explicitly converted to the original objective in equation 2, suggesting RealnessGAN is a generalized version of the original GAN.

- In this paper, analysis concerns the space of probability density functions, where D and G are assumed to have infinite capacities, We start from **finding the optimal realness discriminator D for any given generator G.**
  Then finding the optimal G given optimal D

Theorem 1

Theorem: When G is fixed, for any outcome u and input sample x, the optimal discriminator D satisfies

$$D_G^\star(\boldsymbol{x}, u) = \frac{\mathcal{A}_1(u)p_{data}(\boldsymbol{x}) + \mathcal{A}_0(u)p_g(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}.$$

Proof: Given a fixed G, the objective of D is

$$\min_D V(G, D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(\boldsymbol{x}))],$$

Theorem 1 - Proof

$$\min_D V(G,D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(\boldsymbol{x}))], \tag{5}$$

$$= \int_{\boldsymbol{x}} \left(p_{\text{data}}(\boldsymbol{x}) \int_u \mathcal{A}_1(u) \log \frac{\mathcal{A}_1(u)}{D(\boldsymbol{x},u)} du + p_g(\boldsymbol{x}) \int_u \mathcal{A}_0(u) \log \frac{\mathcal{A}_0(u)}{D(\boldsymbol{x},u)} du \right) dx, \tag{6}$$

$$= - \int_{\boldsymbol{x}} (p_{\text{data}}(\boldsymbol{x}) h(\mathcal{A}_1) + p_g(\boldsymbol{x}) h(\mathcal{A}_0))\, dx$$
$$- \int_{\boldsymbol{x}} \int_u (p_{\text{data}}(\boldsymbol{x}) \mathcal{A}_1(u) + p_g(\boldsymbol{x}) \mathcal{A}_0(u)) \log D(\boldsymbol{x},u) du dx, \tag{7}$$

Where $h(A_1)$ and $h(A_0)$ are their entropies. Marking the first term in equation 7 as C1 since it is irrelevant to D, the objective thus is equivalent to:

Theorem 1 - Proof

Marking the first term in the lase eq as C1 since it is irrelevant to D, the objective thus is equivalent to:

$$\min_D V(G,D) = -\int_{\boldsymbol{x}} (p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})) \int_u \frac{p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \log D(\boldsymbol{x},u)\,du\,dx + C_1$$

Let, $p_{\boldsymbol{x}}(u) = \frac{p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u)+p_g(\boldsymbol{x})\mathcal{A}_0(u)}{p_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x})}$ and $C_2 = p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})$

Then:

$$\begin{aligned}\min_D V(G,D) &= C_1 + \int_{\boldsymbol{x}} C_2 \left(-\int_u p_{\boldsymbol{x}}(u) \log D(\boldsymbol{x},u)\,du + h(p_{\boldsymbol{x}}) - h(p_{\boldsymbol{x}}) \right) dx, \\ &= C_1 + \int_{\boldsymbol{x}} C_2 \mathcal{D}_{\text{KL}}(p_{\boldsymbol{x}} \| D(\boldsymbol{x}))\,dx + \int_{\boldsymbol{x}} C_2 h(p_{\boldsymbol{x}})\,dx.\end{aligned}$$

Observing the last equation, one can see that for any valid x, when $D_{KL}(P_x \,||\, D(x))$ achieves its minimum, D obtains its optimal D^*, leading to $D^*(x) = P_x$ which concludes the proof.

Theorem 2

Next, we move on to the conditions for G to reach its optimal when $D = D^*_G$.

Theorem: When $D = D^*_G$, and there exists an outcome $u \in \Omega$ such that $A_1(u) \neq A_0(u)$, the maximum of $V(G, D^*_G)$ is achieved if and only if $P_g = P_{data}$.

Proof: Our goal is to maximize $V(G, D^*_G)$

First, calculate $V^*(G, D^*_G)$ - the case which $P_g = P_{data}$

$$D_G^\star(\boldsymbol{x}, u) = \frac{\mathcal{A}_1(u)p_{data}(\boldsymbol{x}) + \mathcal{A}_0(u)p_g(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}.$$

Turns into

$$D_G^\star(\boldsymbol{x}, u) = \frac{\mathcal{A}_1(u)+\mathcal{A}_0(u)}{2}$$

Lets, plug $D^*_G(x,u)$ into

$$V(G, D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(\boldsymbol{x}))].$$

Theorem 2 - Proof

Proof:

$$V^\star(G, D_G^\star) = \int_u \mathcal{A}_1(u) \log \frac{2\mathcal{A}_1(u)}{\mathcal{A}_1(u) + \mathcal{A}_0(u)} + \mathcal{A}_0(u) \log \frac{2\mathcal{A}_0(u)}{\mathcal{A}_1(u) + \mathcal{A}_0(u)} du.$$

Now, Subtracting $V^*(G, D^*_G)$ from $V(G, D^*_G)$ gives:

$$\begin{aligned} V'(G, D_G^\star) &= V(G, D_G^\star) - V^\star(G, D_G^\star) \\ &= \int_{\boldsymbol{x}} \int_u (p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)) \log \frac{(p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x}))(\mathcal{A}_1(u) + \mathcal{A}_0(u))}{2(p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u))} du dx, && (12) \\ &= -2 \int_{\boldsymbol{x}} \int_u \frac{p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)}{2} \log \frac{\frac{p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)}{2}}{\frac{(p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x}))(\mathcal{A}_1(u) + \mathcal{A}_0(u))}{4}} du dx, && (13) \\ &= -2\mathcal{D}_{\text{KL}}\left(\frac{p_{\text{data}}\mathcal{A}_1 + p_g\mathcal{A}_0}{2} \,\Big\|\, \frac{(p_{\text{data}} + p_g)(\mathcal{A}_1 + \mathcal{A}_0)}{4}\right). && (14) \end{aligned}$$

Theorem 2 - Proof

Proof:

Since $V^*(G, D^*_G)$ is a constant with respect to G, maximizing $V(G, D^*_G)$ is equivalent to maximizing $V'(G, D^*_G)$. The optimal $V'(G, D^*_G)$ is achieved if and only if the KL divergence reaches its minimum, where:

$$\frac{p_{\text{data}}\mathcal{A}_1 + p_g\mathcal{A}_0}{2} = \frac{(p_{\text{data}} + p_g)(\mathcal{A}_1 + \mathcal{A}_0)}{4}$$

$$(p_{\text{data}} - p_g)(\mathcal{A}_1 - \mathcal{A}_0) = 0.$$

for any valid x and u. Hence, as long as there exists a valid u that $A_1(u) \neq A_0(u)$,
We have $P_{data} = P_g$ for any valid x.

Discussion

- **Effectiveness of anchors:**

  view the last equation as a cost function to minimize, when $P_{data} \neq P_g$, for some $u \in \Omega$, the larger the difference between $A_1(u)$ and $A_0(u)$ is, the stronger the constraint on G becomes. Intuitively, RealnessGAN can be more efficiently trained if we choose A_0 and A_1 to be adequately different.

- **Objective of G:** $\max_G \min_D V(G, D) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x} \sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(\boldsymbol{x}))].$

  according to this equation, the best way to fool D is to increase the KL divergence between D(x) and the anchor distribution A_0 of fake samples, rather than decreasing the KL divergence between D(x) and the anchor distribution A_1 of real samples

Discussion

- **Number of outcomes:**

  In the case of discrete distributions, along with the increment of the number of outcomes, the constraints imposed on G accordingly become more rigorous and can cost G more effort to learn. This is due to the fact that having more outcomes suggests a more fine-grained shape of the realness distribution for G to match.

- **Flexibility of RealnessGAN:**

  As a generalization of the standard framework, it is straightforward to integrate RealnessGAN with different GAN architectures.

Implementation

- The realness distribution prealness chose to be a discrete distribution over N outcomes $\Omega = \{u_0, u_1, \ldots, u_{N-1}\}$. Given an input sample x, the discriminator D returns N probabilities on these outcomes, following:

$$p_{\text{realness}}(\boldsymbol{x}, u_i) = \frac{e^{\boldsymbol{\psi}_i(\boldsymbol{x})}}{\sum_j e^{\boldsymbol{\psi}_j(\boldsymbol{x})}}$$

  Similarly, A_1 and A_0 are discrete distributions defined on Ω.

- As shown in the theoretical analysis, the ideal objective for G is:

$$(G_{\text{objective1}}) \quad \min_G -\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(G(\boldsymbol{z}))]$$

Implementation - G objective

- The discriminator D is not always at its optimal, standard objective in practice could only lead to a generator with limited generative power.
- There are several choices for the regularizer:
  - term that minimizes the KL divergence between D(x) of generated samples and random real samples
  - term that minimizes the KL divergence between A_1 and D(x) of generated samples

$$(G_{\text{objective1}}) \quad \min_G -\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(G(\boldsymbol{z}))]$$

$$(G_{\text{objective2}}) \quad \min_G \quad \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}},\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(D(\boldsymbol{x}) \| D(G(\boldsymbol{z}))] - \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(G(\boldsymbol{z}))],$$

$$(G_{\text{objective3}}) \quad \min_G \quad \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(G(\boldsymbol{z}))] - \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(G(\boldsymbol{z}))].$$

Implementation - Feature resampling

Introduce a resampling technique performed on the realness output to augment data variance

Given a mini-batch $\{x_0, \ldots, x_{M-1}\}$, a Gaussian distribution $N(\mu_i, \sigma_i)$ is fitted on $\{\psi_i(x_0), \psi_i(x_1), \ldots, \psi_i(x_{M-1})\}$, which are logits computed by D on i-th outcome.
Then, resample M new logits $\{\psi_i'(x_0), \psi_i'(x_1), \ldots, \psi_i'(x_{M-1}); \psi_i' \sim N(\mu_i, \sigma_i)\}$ for i-th outcome and use them succeedingly.

Advantages using resampling technique:

- More robust models.
- Demands instances of ψi(x) to be homologous throughout the mini-batch, such that each outcome reflects realness consistently across samples.

Experiments - Synthetic dataset

- 100,000 2D points sampled from a mixture of 9 isotropic Gaussian distributions whose means are arranged in a 3 by 3 grid, with variances equal to 0.05.

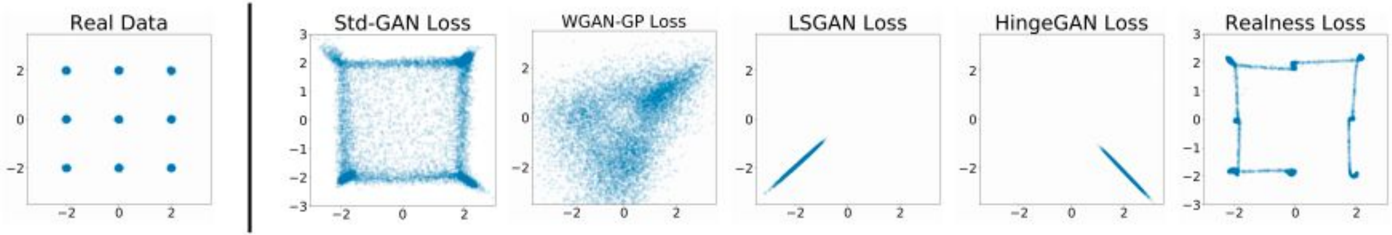


Figure 2: Left: real data sampled from the mixture of 9 Gaussian distributions. Right: samples generated by *Std-GAN*, *WGAN-GP*, *LSGAN*, *HingeGAN* and *RealnessGAN*.

- To evaluate P_g, draw 10,000 samples and measure their quality and diversity

Experiments - Synthetic dataset

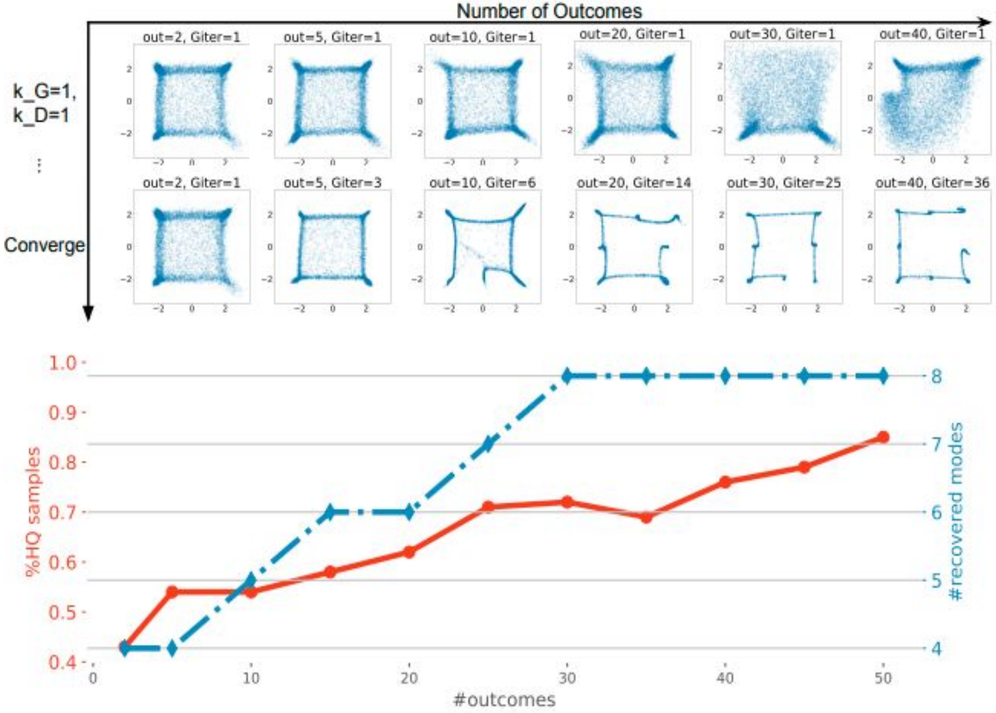
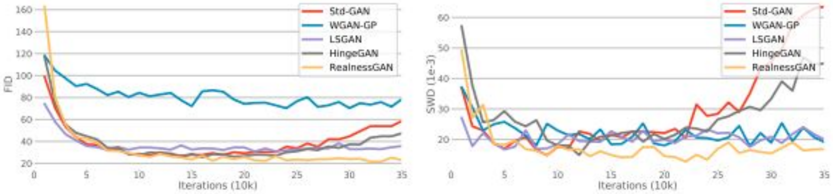


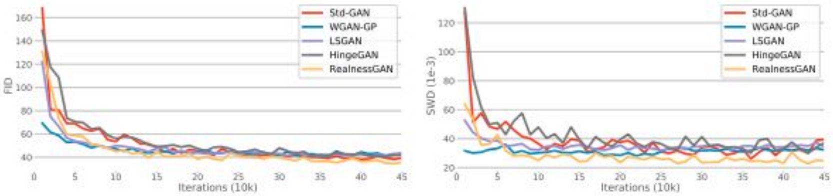
Figure 3: First row: the results of *RealnessGAN* when fixing $k_G = k_D = 1$ and increasing the number of outcomes. Second row: the results of *RealnessGAN* when k_G is properly increased. Bottom curves: under the settings of second row, the ratio of high quality samples and the number of recovered modes.

Experiments - Real World datasets



(a) FID on CelebA

(b) SWD on CelebA



	Method	FID ↓				SWD ($\times 10^3$) ↓			
		Min	Max	Mean	SD	Min	Max	Mean	SD
CelebA	Std-GAN	27.02	70.43	34.85	9.40	14.81	68.06	30.58	15.39
	WGAN-GP	70.28	104.60	81.15	8.27	17.85	30.56	22.09	2.93
	LSGAN	30.76	57.97	34.99	5.15	16.72	23.99	20.39	2.25
	HingeGAN	25.57	75.03	33.89	10.61	14.91	54.30	28.86	10.34
	RealnessGAN	**23.51**	81.3	**30.82**	7.61	**12.72**	31.39	**17.11**	3.59
CIFAR10	Std-GAN	38.56	88.68	47.46	15.96	28.76	57.71	37.55	7.02
	WGAN-GP	41.86	79.25	46.96	5.57	28.17	36.04	30.98	1.78
	LSGAN	42.01	75.06	48.41	7.72	31.99	40.46	34.75	2.34
	HingeGAN	42.40	117.49	57.30	20.69	32.18	61.74	41.85	7.31
	RealnessGAN	**34.59**	102.98	**42.30**	11.84	**22.80**	53.38	**26.98**	5.47

Experiments - Resampling technique

Despite the results are similar,
feature resampling stabilizes the training process especially in the latter stage.

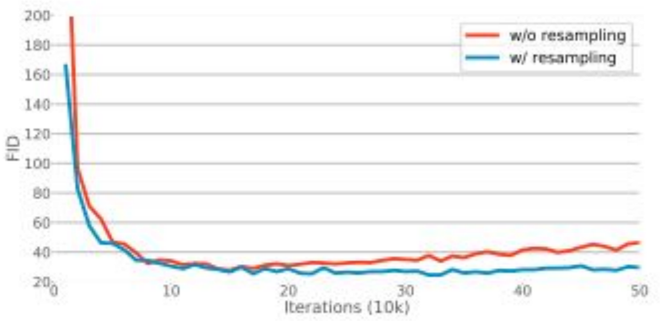


Figure 5: Training FID curves of *Realness-GAN* with and without feature re-sampling.

Experiments - Different anchor distributions

$\mathcal{D}_{\mathrm{KL}}(\mathcal{A}_1 \| \mathcal{A}_0)$	Min	Max	Mean	SD
1.66	31.01	96.11	40.75	11.83
5.11	26.22	87.98	36.11	9.83
7.81	25.98	85.51	36.30	10.04
11.05	23.51	81.30	30.82	7.61

Figure 6: Samples generated by *RealnessGAN* trained with the ideal objective (equation 18). Top-row: samples when $\mathcal{D}_{\mathrm{KL}}(\mathcal{A}_1 \| \mathcal{A}_0) = 11.05$. Bottom-row: samples when $\mathcal{D}_{\mathrm{KL}}(\mathcal{A}_1 \| \mathcal{A}_0) = 33.88$.

Experiments - Different G objectives

Table 3: FID scores of G on CIFAR10, trained with different objectives.

G Objective	FID
Objective1 (equation 18)	36.73
Objective2 (equation 19)	34.59
Objective3 (equation 20)	36.21
DCGAN	38.56
WGAN-GP	41.86
LSGAN	42.01
HingeGAN	42.40

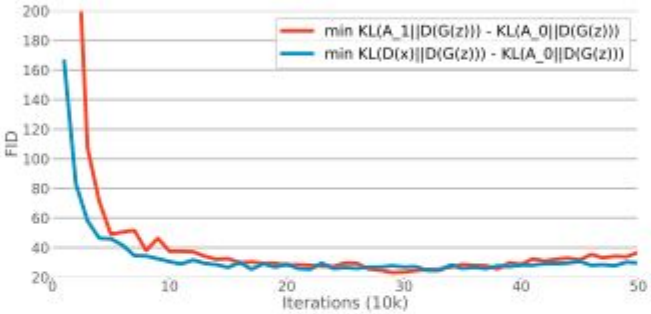


Figure 7: Training curves of *RealnessGAN* on CelebA using objective2 (equation 19) and objective3 (equation 20).

Questions ?