Real or not real, that is the question

(ICLR 2020 spotlight)

The Chinese University of Hong Kong

Observation

- What makes those pictures look real or not?
    - Inharmonious facial structure and components
    - Unnatural background
    - Abnormal style combination
    - Texture distortion
    - ...

Observation

- Discriminator predict the picture is real or not
  - Although there are several perspectives for determining, the output from the discriminator is a single scalar
  - The single scalar could be viewed as an abstract or summarization of multiple measures

Idea

- RealnessGAN
  - The output of discriminator is a distribution
  - Standard GAN can be viewed as a special case of RealnessGAN

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

$$= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log(\underline{D(\boldsymbol{x}) - 0})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log(\underline{1 - D(\boldsymbol{x})})], \quad (2)$$

  - Interpretation: difference between
    one dimensional vector D(x) and one dimensional vector [0] / [1]

Idea

- RealnessGAN

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

$$= \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log(\underline{D(\boldsymbol{x}) - 0})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log(\underline{1 - D(\boldsymbol{x})})], \quad (2)$$

  - Interpretation: difference between one dimensional vector D(x) and one dimensional vector [0] / [1]

  - What if a multi-dimensional case?

$$\max_G \min_D V(G, D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(\boldsymbol{x}))]. \quad (3)$$

  - Difference between two multi-dimensional vectors (distrib.)

Theory

Future works:
Replacing KL divergence by earth mover distance
Applying on improved architectures (progressiveGAN, styleGAN)

- Ground-truth distribution
  - A1: real, A0: fake (we call these anchor distributions)

$$\max_G \min_D V(G, D) = \mathbb{E}_{\boldsymbol{x} \sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x} \sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(\boldsymbol{x}))]. \quad (3)$$

- Optimality $\boxed{D_G^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \text{ for a fixed } G}$

Theorem 1. *When G is fixed, for any outcome u and input sample $\boldsymbol{x}$, the optimal discriminator D satisfies*

$$D_G^\star(\boldsymbol{x}, u) = \frac{\mathcal{A}_1(u) p_{data}(\boldsymbol{x}) + \mathcal{A}_0(u) p_g(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}. \quad (4)$$

Theory

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

$$\min_D V(G,D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1\|D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0\|D(\boldsymbol{x}))], \tag{17}$$

$$= \int_{\boldsymbol{x}} \left(p_{\text{data}}(\boldsymbol{x}) \int_u \mathcal{A}_1(u) \log \frac{\mathcal{A}_1(u)}{D(\boldsymbol{x},u)} du + p_g(\boldsymbol{x}) \int_u \mathcal{A}_0(u) \log \frac{\mathcal{A}_0(u)}{D(\boldsymbol{x},u)} du \right) dx, \tag{18}$$

$$= -\int_{\boldsymbol{x}} \left(p_{\text{data}}(\boldsymbol{x}) h(\mathcal{A}_1) + p_g(\boldsymbol{x}) h(\mathcal{A}_0)\right) dx$$
$$- \int_{\boldsymbol{x}} \int_u \left(p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)\right) \log D(\boldsymbol{x},u) du dx, \tag{19}$$

Theory

$$
\begin{aligned}
&= -\int_{\boldsymbol{x}} \left(p_{\text{data}}(\boldsymbol{x})h(\mathcal{A}_1) + p_g(\boldsymbol{x})h(\mathcal{A}_0)\right) dx \\
&- \int_{\boldsymbol{x}}\int_{u} \left(p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)\right) \log D(\boldsymbol{x}, u) du dx, \qquad (19)
\end{aligned}
$$

where $h(\mathcal{A}_1)$ and $h(\mathcal{A}_0)$ are their entropies, and the first term in equation 19 is irrelevant to D, marked as C_1. The objective thus is equivalent to:

$$
\min_{D} V(G, D) = -\int_{\boldsymbol{x}}\int_{u} \left(p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)\right) \log D(\boldsymbol{x}, u) du dx + C_1, \qquad (20)
$$

$$
= -\int_{\boldsymbol{x}} \left(p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})\right) \int_{u} \frac{p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \log D(\boldsymbol{x}, u) du dx + C_1, \qquad (21)
$$

where $p_{\boldsymbol{x}}(u) = \frac{p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}$ is a distribution defined on Ω_u. Consequently, let $C_2 = p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})$, we have

$$
\min_{D} V(G, D) = C_1 + \int_{\boldsymbol{x}} C_2 \left(-\int_{u} p_{\boldsymbol{x}}(u) \log D(\boldsymbol{x}, u) du + h(p_{\boldsymbol{x}}) - h(p_{\boldsymbol{x}})\right) dx, \qquad (22)
$$

$$
= C_1 + \int_{\boldsymbol{x}} C_2 \mathcal{D}_{\text{KL}}(p_{\boldsymbol{x}} \| D(\boldsymbol{x})) dx + \int_{\boldsymbol{x}} C_2 h(p_{\boldsymbol{x}}) dx. \qquad (23)
$$

From equation 23 we can see, for any $\boldsymbol{x} \in Supp(p_{\text{data}}) \cup Supp(p_g)$, when $\mathcal{D}_{\text{KL}}(p_{\boldsymbol{x}} \| D(\boldsymbol{x}))$ achieves its minimum, D obtains its optimal $D^\star$. And at that time, we have $D^\star(\boldsymbol{x}) = p_{\boldsymbol{x}}$, which concludes the proof. $\square$

Theory

$$D_G^\star(\boldsymbol{x}, u) = \frac{\mathcal{A}_1(u)p_{data}(\boldsymbol{x}) + \mathcal{A}_0(u)p_g(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})}. \tag{4}$$

$$\min_D V(G, D) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1 \| D(\boldsymbol{x}))] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0 \| D(\boldsymbol{x}))], \tag{17}$$

$$= \int_{\boldsymbol{x}} \left(p_{\text{data}}(\boldsymbol{x}) \int_u \mathcal{A}_1(u) \log \frac{\mathcal{A}_1(u)}{D(\boldsymbol{x}, u)} du + p_g(\boldsymbol{x}) \int_u \mathcal{A}_0(u) \log \frac{\mathcal{A}_0(u)}{D(\boldsymbol{x}, u)} du \right) dx, \tag{18}$$

Theorem 2. *When $D = D_G^\star$, and there exists an outcome $u \in \Omega$ such that $\mathcal{A}_1(u) \neq \mathcal{A}_0(u)$, the maximum of $V(G, D_G^\star)$ is achieved if and only if $p_g = p_{data}$.*

Proof. When $p_g = p_{\text{data}}$, $D_G^\star(\boldsymbol{x}, u) = \frac{\mathcal{A}_1(u)+\mathcal{A}_0(u)}{2}$, we have:

$$V^\star(G, D_G^\star) = \int_u \mathcal{A}_1(u) \log \frac{2\mathcal{A}_1(u)}{\mathcal{A}_1(u) + \mathcal{A}_0(u)} + \mathcal{A}_0(u) \log \frac{2\mathcal{A}_0(u)}{\mathcal{A}_1(u) + \mathcal{A}_0(u)} du. \tag{7}$$

Subtracting $V^\star(G, D_G^\star)$ from $V(G, D_G^\star)$ gives:

$$V'(G, D_G^\star) = V(G, D_G^\star) - V^\star(G, D_G^\star)$$

$$= \int_{\boldsymbol{x}} \int_u (p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)) \log \frac{(p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x}))(\mathcal{A}_1(u) + \mathcal{A}_0(u))}{2(p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u))} dudx, \tag{8}$$

$$= -2 \int_{\boldsymbol{x}} \int_u \frac{p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u) + p_g(\boldsymbol{x})\mathcal{A}_0(u)}{2} \log \frac{\frac{p_{\text{data}}(\boldsymbol{x})\mathcal{A}_1(u)+p_g(\boldsymbol{x})\mathcal{A}_0(u)}{2}}{\frac{(p_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x}))(\mathcal{A}_1(u)+\mathcal{A}_0(u))}{4}} dudx, \tag{9}$$

$$= -2\mathcal{D}_{\text{KL}}\left(\frac{p_{\text{data}}\mathcal{A}_1 + p_g\mathcal{A}_0}{2} \Big\| \frac{(p_{\text{data}} + p_g)(\mathcal{A}_1 + \mathcal{A}_0)}{4}\right). \tag{10}$$

Theory

$$= -2\mathcal{D}_{\mathrm{KL}}\left(\frac{p_{\mathrm{data}}\mathcal{A}_1 + p_g\mathcal{A}_0}{2} \middle\| \frac{(p_{\mathrm{data}} + p_g)(\mathcal{A}_1 + \mathcal{A}_0)}{4}\right). \tag{10}$$

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

$$\frac{p_{\mathrm{data}}\mathcal{A}_1 + p_g\mathcal{A}_0}{2} = \frac{(p_{\mathrm{data}} + p_g)(\mathcal{A}_1 + \mathcal{A}_0)}{4}, \tag{11}$$

$$(p_{\mathrm{data}} - p_g)(\mathcal{A}_1 - \mathcal{A}_0) = 0, \tag{12}$$

for any valid $\boldsymbol{x}$ and u. Hence, as long as there exists a valid u that $\mathcal{A}_1(u) \neq \mathcal{A}_0(u)$, we have $p_{\mathrm{data}} = p_g$ for any valid $\boldsymbol{x}$. $\square$

Discussion

- Number of outcomes
  - Increment of the number of outcomes makes G become rigorous and G needs more effort to learn
  - It is related to the ratio of the number of updates between G and D

Discussion

- Objective of G

As shown in the theoretical analysis, the ideal objective for G is maximizing the KL divergence between $D(\boldsymbol{x})$ of generated samples and $\mathcal{A}_0$:

$$(G_{\text{objective1}}) \qquad \min_G -\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0\|D(G(\boldsymbol{z}))]. \tag{14}$$

  - Alternative objectives (regularization)
    - Since D is not always optimal (A1: real, A0: fake)

$$(G_{\text{objective2}}) \qquad \min_G \quad \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}},\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(D(\boldsymbol{x})\|D(G(\boldsymbol{z}))] - \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0\|D(G(\boldsymbol{z}))], \tag{15}$$

$$(G_{\text{objective3}}) \qquad \min_G \quad \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_1\|D(G(\boldsymbol{z}))] - \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\text{KL}}(\mathcal{A}_0\|D(G(\boldsymbol{z}))]. \tag{16}$$

Experiment

- Synthetic dataset
  - Real distribution: the mixture of nine normal distributions
  - Generator and discriminator are consists or four FC layers
  - Input latent from 32 dimensional normal distribution

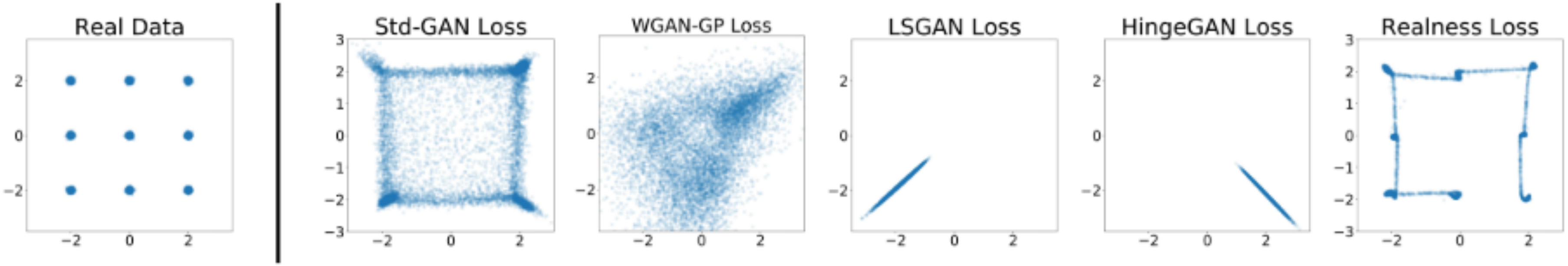
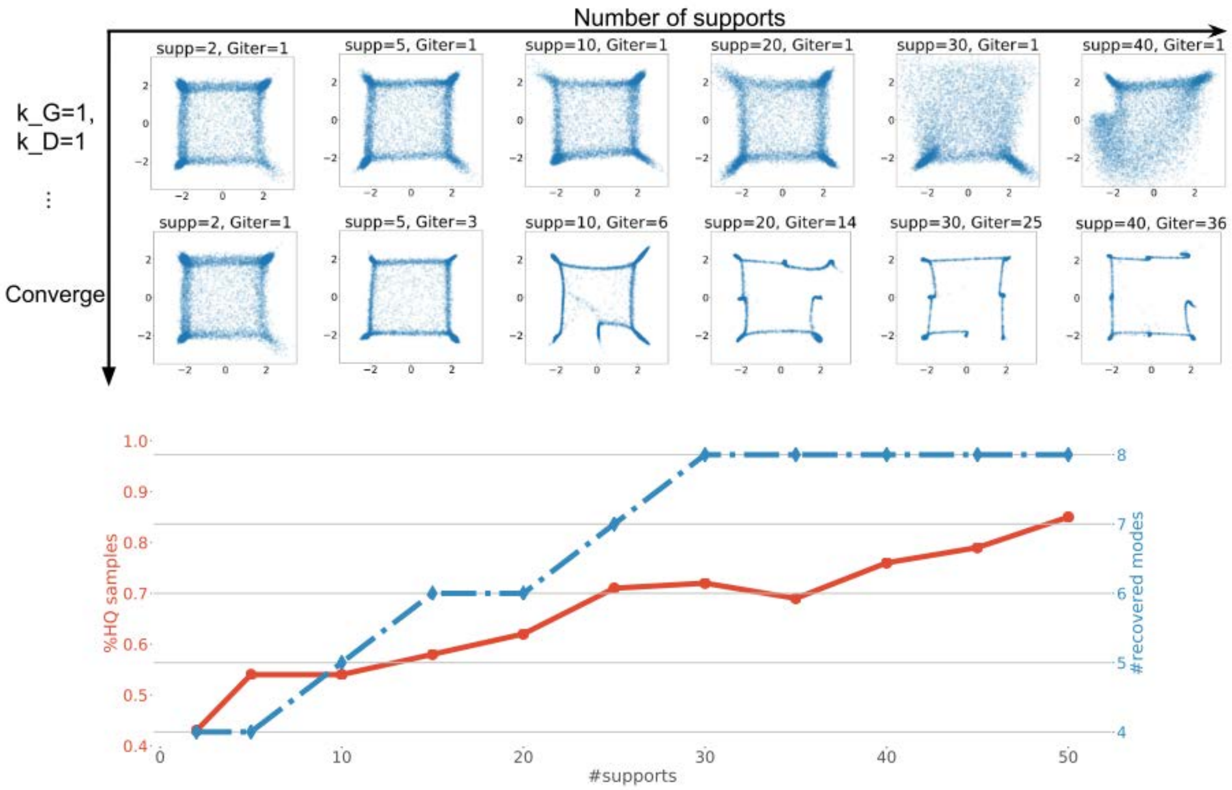


Figure 2: Left: real data sampled from the mixture of 9 Gaussian distributions. Right: samples generated by *Std-GAN*, *WGAN-GP*, *LSGAN*, *HingeGAN* and *RealnessGAN*.

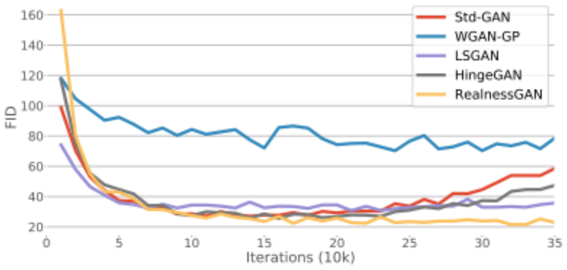
Experiments

- Synthetic dataset
  - Effect of adjusting the number of outcomes (supports)

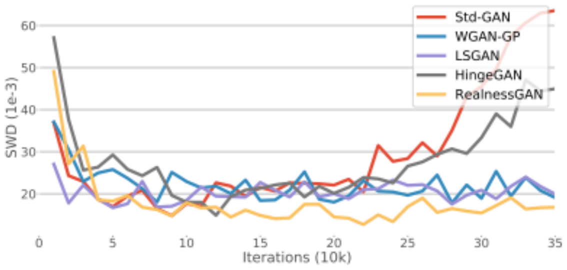


Experiments

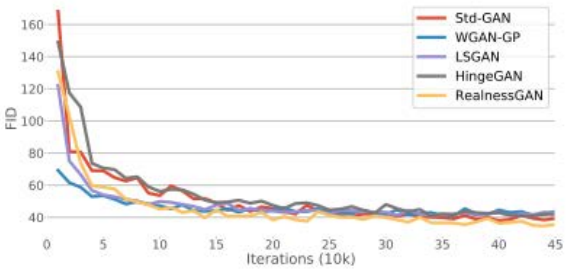
- Real-world datasets
  - CelebA, CIFAR-10 and FFHQ
    (32 by 32, 256 by 256 and 1,024 by 1,024)
  - Use DCGAN



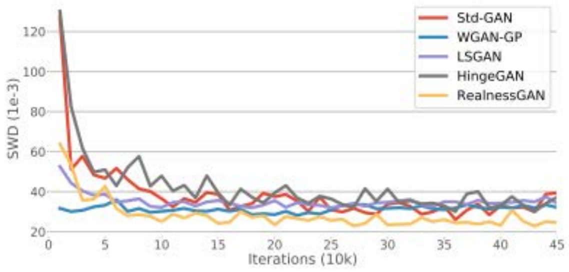
(a) FID on CelebA



(b) SWD on CelebA



(c) FID on CIFAR10



(d) SWD on CIFAR10

Experiments

- Objective of G

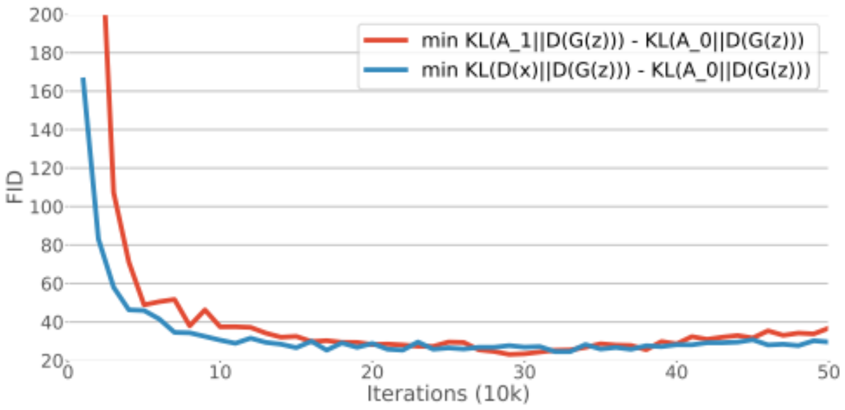
$$(G_{\text{objective1}}) \qquad \min_G -\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\mathrm{KL}}(\mathcal{A}_0 \| D(G(\boldsymbol{z}))]. \tag{14}$$

$$(G_{\text{objective2}}) \qquad \min_G \quad \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}},\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\mathrm{KL}}(D(\boldsymbol{x}) \| D(G(\boldsymbol{z}))] - \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\mathrm{KL}}(\mathcal{A}_0 \| D(G(\boldsymbol{z}))], \tag{15}$$

$$(G_{\text{objective3}}) \qquad \min_G \quad \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\mathrm{KL}}(\mathcal{A}_1 \| D(G(\boldsymbol{z}))] - \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\mathcal{D}_{\mathrm{KL}}(\mathcal{A}_0 \| D(G(\boldsymbol{z}))]. \tag{16}$$

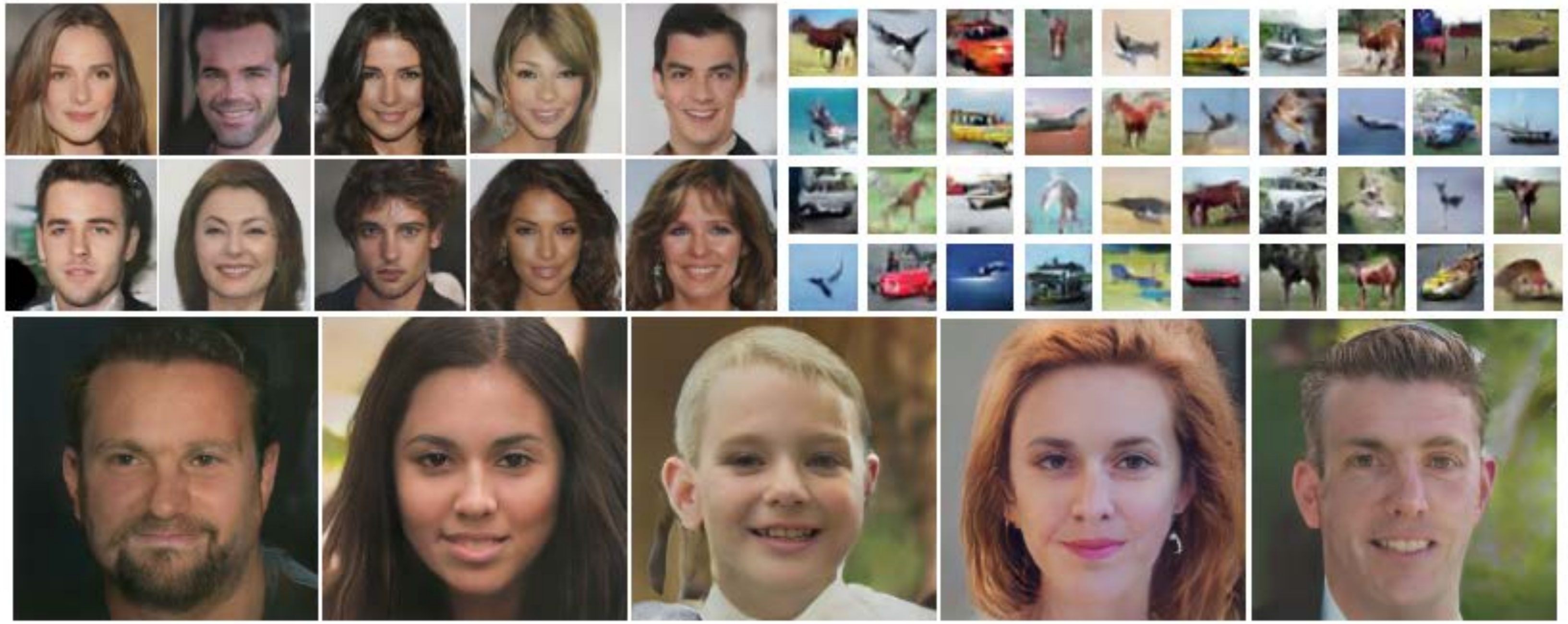
Table 3: In this table we compare different objectives of G on CIFAR10.

Objective	FID
Ideal Objective (equation 14)	36.73
Objective 2 (equation 15)	34.59
Objective 3 (equation 16)	36.21



Experiments

- Real-world datasets
  - 1024 by 1024 images from DCGAN with RealnessGAN



Experiments

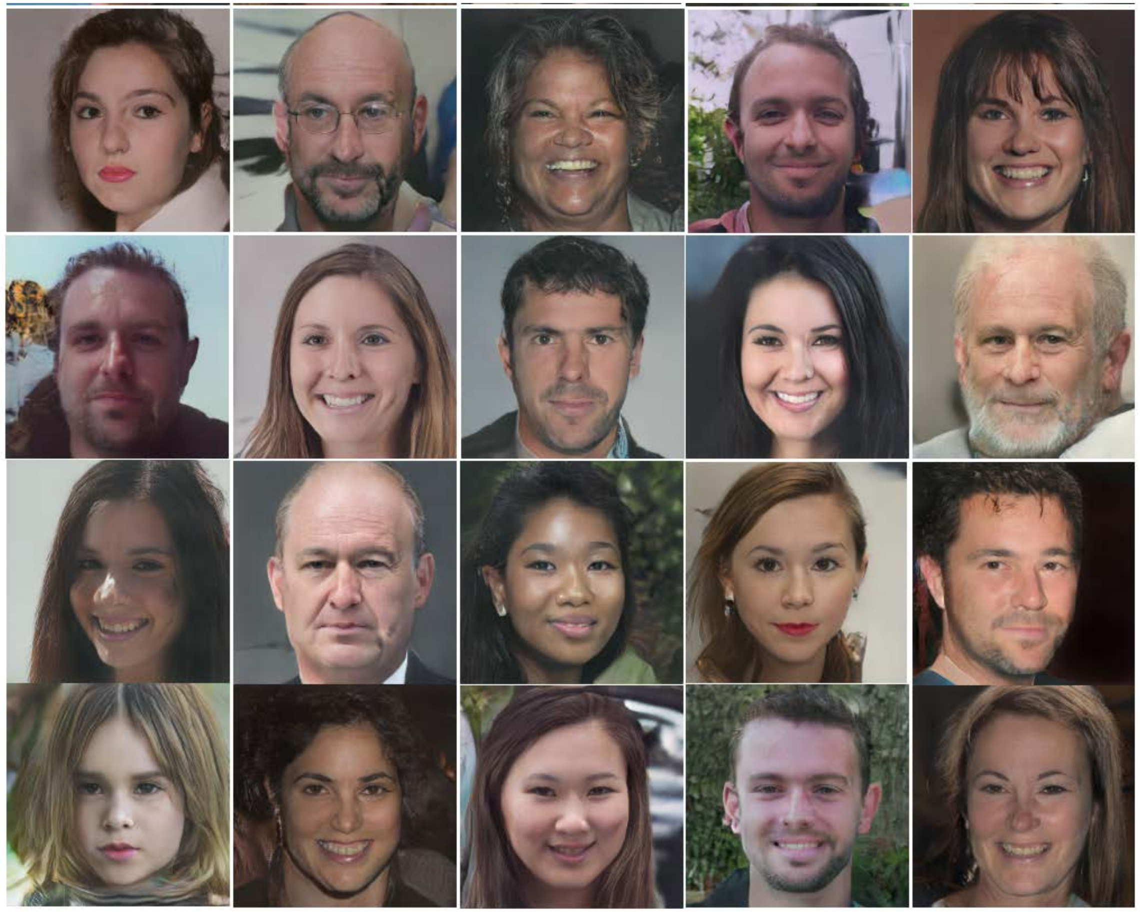
We further compute FID as the quantitative result. Specifically, RealnessGAN yields a FID score of *17.18*. For reference, we also re-implement StyleGAN and train it using the same setting, resulting in a FID score of *16.12*.

- Real-world datasets
  - 1024 by 1024 images from DCGAN with RealnessGAN

Table 1: Minimum (min), maximum (max), mean and standard deviation (SD) of FID and SWD on CelebA and CIFAR10, calculated at 20k, 30k, ... iterations. The best indicators in baseline methods are underlined.

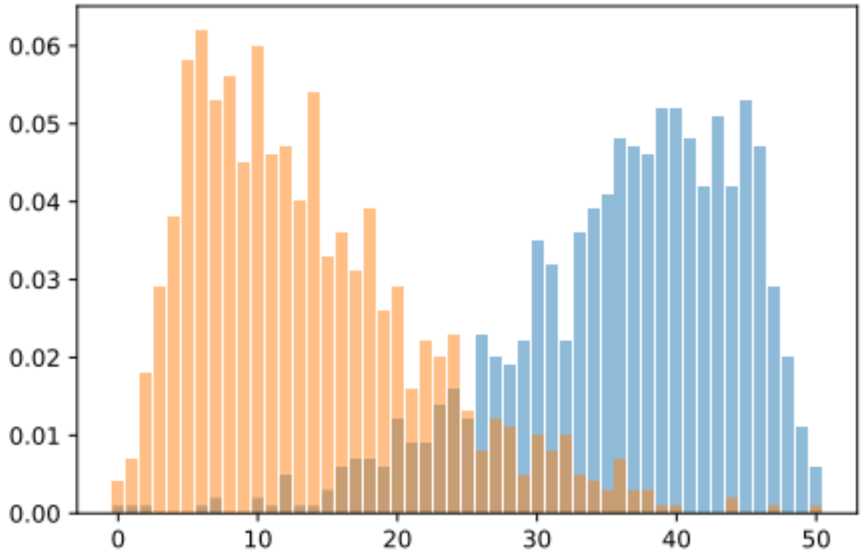
	Method	**FID** $\downarrow$				**SWD** $(\times 10^3)\downarrow$			
		Min	Max	Mean	SD	Min	Max	Mean	SD
CelebA	Std-GAN	27.02	70.43	34.85	9.40	14.81	68.06	30.58	15.39
	WGAN-GP	70.28	104.60	81.15	8.27	17.85	30.56	22.09	2.93
	LSGAN	30.76	57.97	34.99	5.15	16.72	23.99	20.39	2.25
	HingeGAN	25.57	75.03	33.89	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



Issues

- Selecting anchor distribution
  - In this paper, the author chose two skewed normal distribution



Issues

- Multi-dimensional output VS multiple discriminator
  - Is it same as an ensemble of discriminators? (Ensemble GAN)
  - No!

  - Conceptually, RealnessGAN and EnsembledGAN are orthogonal
  - RealnessGAN could serve as one of the discriminators of EnsembleGAN
  - Technically, EnsembleGAN uses multiple discriminators that could have different architectures and weights
  - RealnessGAN uses a single discriminator
  - EnsembleGAN using DCGAN fails on FFHQ

Issues

- Role of each outcomes
  - 리버탈에 서술했지만, 아무 말 대단치
  - Future work

Etc

- Code: will be uploaded
- Content in magenta
- Review score: 8, 3->6, 3->6