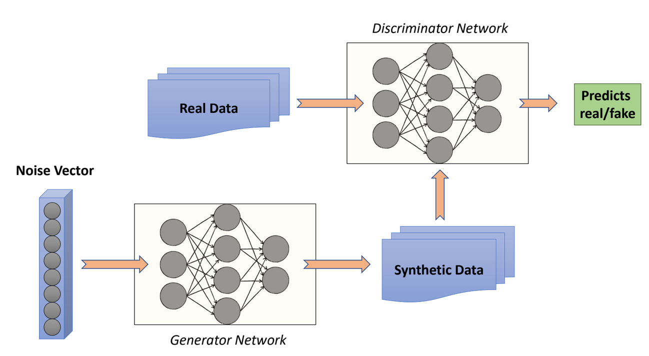
Tunable Dual-Objective GANs for Stable Training

Monica Welfert, Kyle Otstot, Gowtham R. Kurri, and Lalitha Sankar

Arizona State University

Generative Adversarial Networks (GANs)

- GANs [1] are generative models that learn to produce new samples from an unknown (real) distribution P_r
- Generator G_θ and discriminator D_ω play an adversarial game
- G_θ maps noise Z to synthetic samples X_g to mimic the real samples X_r, while D_ω tries to differentiate between the synthetic and real samples
- Formulated as a zero-sum min-max game: $\boxed{\inf_{G_\theta} \sup_{D_\omega} V(\theta, \omega)}$



Various Value Functions & GANs

- Vanilla GAN (Goodfellow et al. [1]) minimizes Jensen-Shannon divergence (JSD):

$$\inf_{G_\theta} \underbrace{\sup_{D_\omega:\mathcal{X}\to[0,1]} \mathbb{E}_{X_r\sim P_r}[\log D_\omega(X_r)] + \mathbb{E}_{X_g\sim P_{G_\theta}}[\log(1-D_\omega(X_g))]}_{=2\mathrm{JSD}(P_r\|P_{G_\theta})-\log 4}$$

- Can reformulate GANs using class probability estimation (CPE) loss $\ell(y,\hat{y})$, $(y,\hat{y}) \in \{0,1\}\times[0,1]$ [2, 3] as

$$\inf_{G_\theta} \sup_{D_\omega:\mathcal{X}\to[0,1]} \boxed{(V_\ell(\theta,\omega) := \mathbb{E}_{X_r\sim P_r}[-\ell(1,D_\omega(X_r))] + \mathbb{E}_{X_g\sim P_{G_\theta}}[-\ell(0,D_\omega(X_g))])}$$

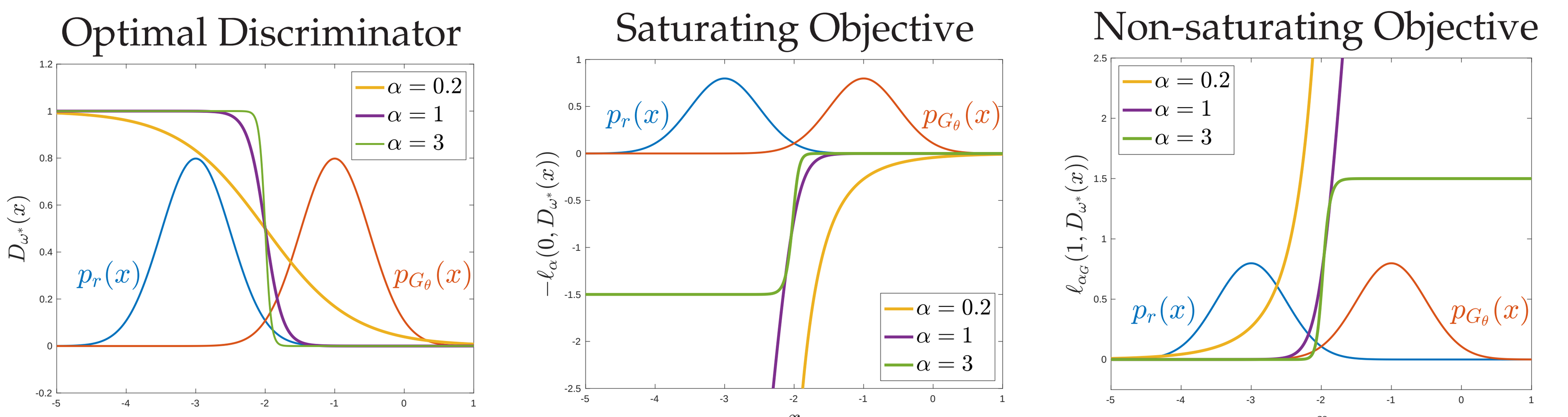
- We obtain α-GAN using α-loss (Sypherd et al. [4])

$$\ell_\alpha(y,\hat{y}) = \frac{\alpha}{\alpha-1}\left(1 - y\hat{y}^{\frac{\alpha-1}{\alpha}} - (1-y)(1-\hat{y})^{\frac{\alpha-1}{\alpha}}\right), \quad \text{for } \alpha \in (0,1)\cup(1,\infty)$$

- α-GAN minimizes the Arimoto divergence and recovers vanilla GAN ($\alpha \to 1$), Hellinger GAN ($\alpha = 1/2$), and total variation (TV) GAN ($\alpha \to \infty$)

Training Instabilities in GANs

Toy example: $P_r = \mathcal{N}(-3, 0.5)$, $P_{G_\theta} = \mathcal{N}(-1, 0.5)$



$$D_{\omega^*}(x) = \frac{p_r(x)^\alpha}{p_r(x)^\alpha + p_{G_\theta}(x)^\alpha} \qquad \ell_\alpha(0,\hat{y}) = \frac{\alpha}{\alpha-1}\left(1-(1-\hat{y})^{\frac{\alpha-1}{\alpha}}\right) \qquad \ell_\alpha(1,\hat{y}) = \frac{\alpha}{\alpha-1}\left(1-\hat{y}^{\frac{\alpha-1}{\alpha}}\right)$$

- Vanilla GAN generator's objective can saturate when discriminator confidently classifies generated data as fake; tuning $\alpha < 1$ addresses *vanishing gradients* by reducing confidence of discriminator
- However, $\alpha \leq 1$ can produce *exploding gradients* for the generator as the generated samples approach real samples, potentially resulting in the generated data being repelled from the real data
- [1] proposed a *non-saturating* alternative generator objective to combat vanishing gradients:

$$\mathbb{E}_{X_g\sim P_{G_\theta}}[-\log(1-D_\omega(X_g))]$$

  - However, this objective can still lead to *model oscillation* and even *mode collapse* due to failure to converge and sensitivity to hyperparameter initialization (e.g. learning rate) because of large gradients
- Can address all of these types of instabilities via different α values for discriminator and generator losses

(α_D, α_G)-GANs: Dual Objectives

- Saturating (α_D, α_G)-GAN [5] non-zero sum game given by:

$$\boxed{\sup_{D_\omega:\mathcal{X}\to[0,1]} V_{\ell_{\alpha_D}}(\theta,\omega)} \qquad \boxed{\inf_{G_\omega} V_{\ell_{\alpha_G}}(\theta,\omega)}$$

Result. *For a fixed G_ω, the D_{ω^*} of an (α_D, α_G)-GAN is the same as that of α-GAN with $\alpha = \alpha_D$. For this D_{ω^*} and for $(\alpha_D, \alpha_G) \in (0,\infty]^2$ such that $(\alpha_D \leq 1,\ \alpha_G > \alpha_D/(\alpha_D+1))$ or $(\alpha_D > 1,\ \alpha_D/2 < \alpha_G \leq \alpha_D)$, the generator of a saturating (α_D, α_G)-GAN minimizes a non-negative symmetric f-divergence.*

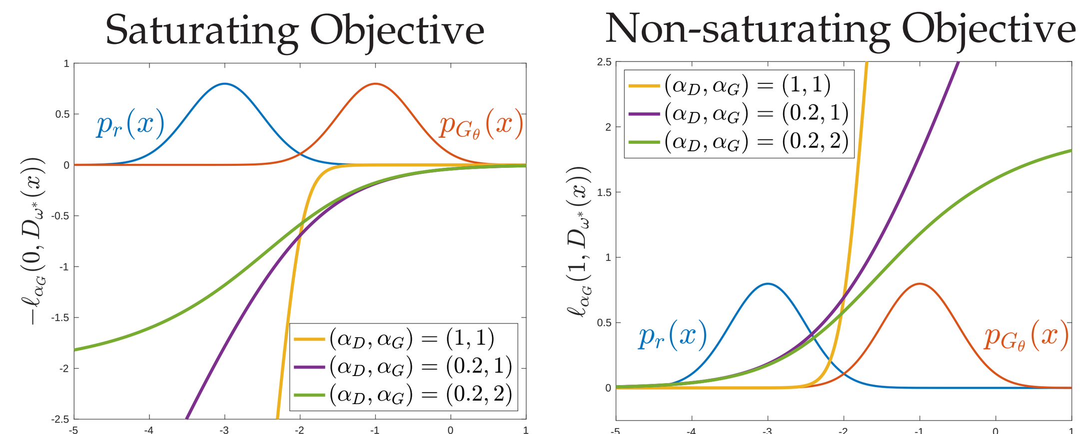
- Non-saturating (α_D, α_G)-GAN given by:

$$\boxed{\sup_{D_\omega:\mathcal{X}\to[0,1]} V_{\ell_{\alpha_D}}(\theta,\omega)} \qquad \boxed{\inf_{G_\omega} \mathbb{E}_{X_g\sim P_{G_\theta}}[\ell_{\alpha_G}(1, D_\omega(X_g))]}$$

Result. *For the same D_{ω^*} and for $(\alpha_D, \alpha_G) \in (0,\infty]^2$ with $\alpha_D + \alpha_G > \alpha_G\alpha_D$, the generator of a non-saturating (α_D, α_G)-GAN minimizes a non-negative asymmetric f-divergence.*

(α_D, α_G)-GANs: Toy Example

Toy example: $P_r = \mathcal{N}(-3, 0.5)$, $P_{G_\theta} = \mathcal{N}(-1, 0.5)$



Tuning $\alpha_D < 1$ and $\alpha_G = 1$ produces more gradient for the generator while making its objective less convex, which helps stabilize training; tuning $\alpha_G > 1$ results in a quasiconvex generator objective, which can further improve training stability

Illustration of Results

- **2D-ring dataset**: samples drawn from a mixture of 8 equal-prior Gaussian distributions (*modes*), indexed $i \in \{1, 2, \ldots, 8\}$ with mean $(\cos(2\pi i/8), \sin(2\pi i/8))$ and variance 10^{-4}

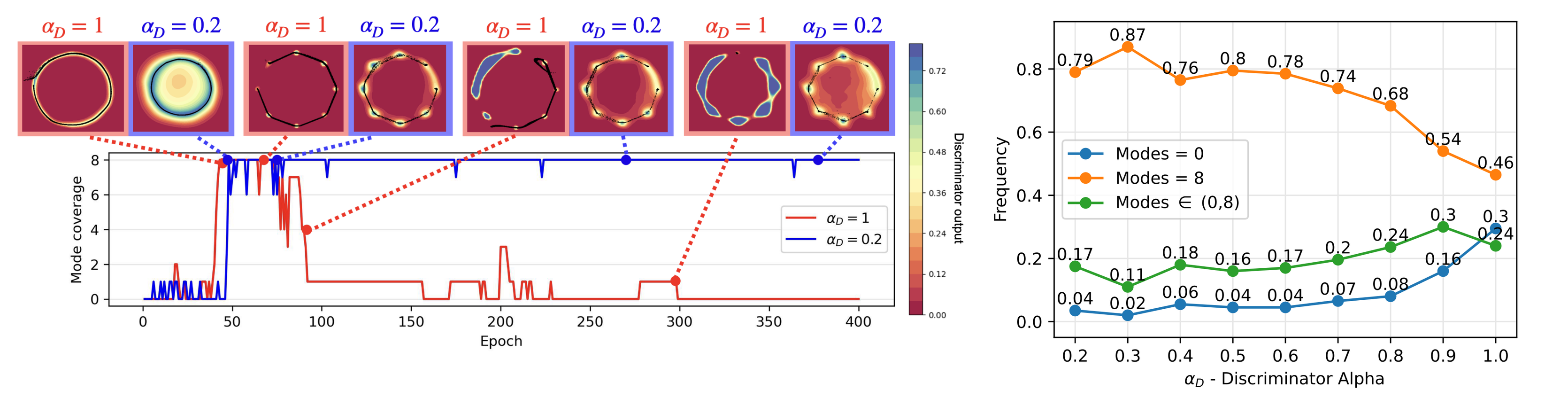


Figure 1: (Left) Plot of mode coverage over epochs for *saturating* (α_D, α_G)-GAN, fixing $\alpha_G = 1$. (Right) Plot of success and failure rates over 200 seeds for a range of α_D values with $\alpha_G = 1$.

- **Celeb-A dataset**: collection of over 200,000 celebrity headshots, resized to 64×64
- Compare performance of non-saturating vanilla GAN, non-saturating (α_D, α_G)-GANs and **Least Squares GAN (LSGAN)** [6] with 0-1 binary coding scheme ($a = 0, b = c = 1$):

$$\text{D: } \inf_{\omega\in\Omega} \mathbb{E}_{X_r\sim P_r}\left[\frac{1}{2}(D_\omega(X_r)-b)^2\right] + \mathbb{E}_{X_g\sim P_{G_\theta}}\left[\frac{1}{2}(D_\omega(X_g)-a)^2\right]$$

$$\text{G: } \inf_{\theta\in\Theta} \mathbb{E}_{X_g\sim P_{G_\theta}}\left[\frac{1}{2}(D_\omega(X_g)-c)^2\right]$$

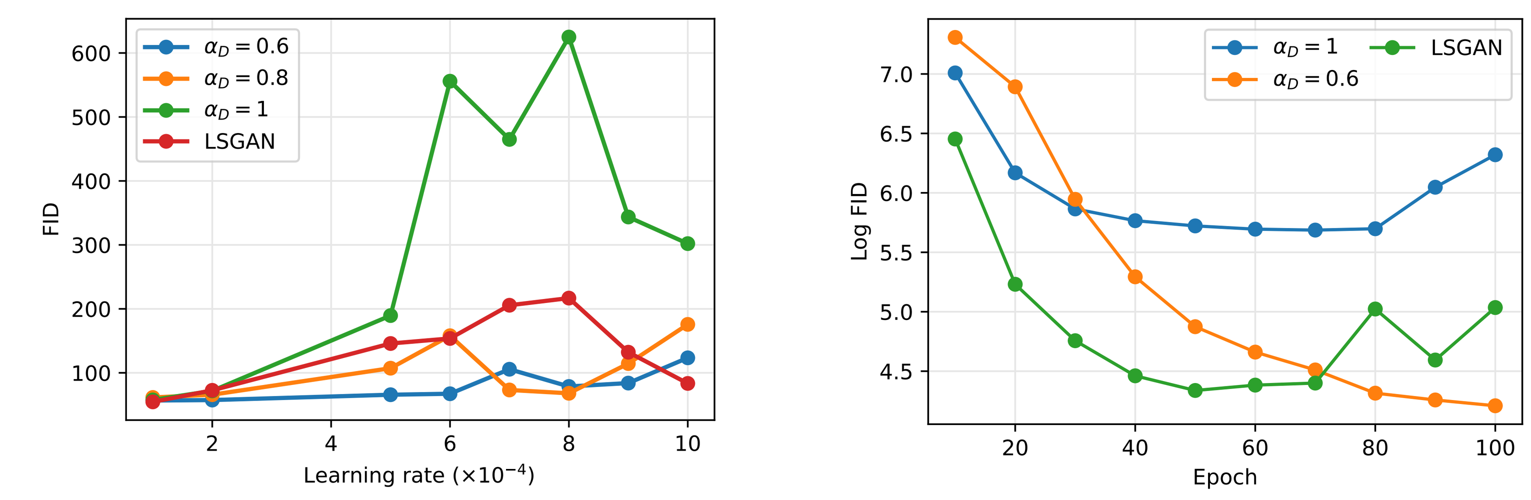


Figure 2: (Left) Plot of FID (smaller is better) averaged over 50 seeds vs. learning rate for a fixed number of epochs (=100) and different non-saturating $(\alpha_D, \alpha_G = 1)$-GANs as well as LSGAN. (Right) Log-scale plot of FID over training epochs for the non-saturating $(1,1)$-GAN (vanilla), the non-saturating $(0.6, 1)$-GAN and LSGAN with learning rate 6×10^{-4}.

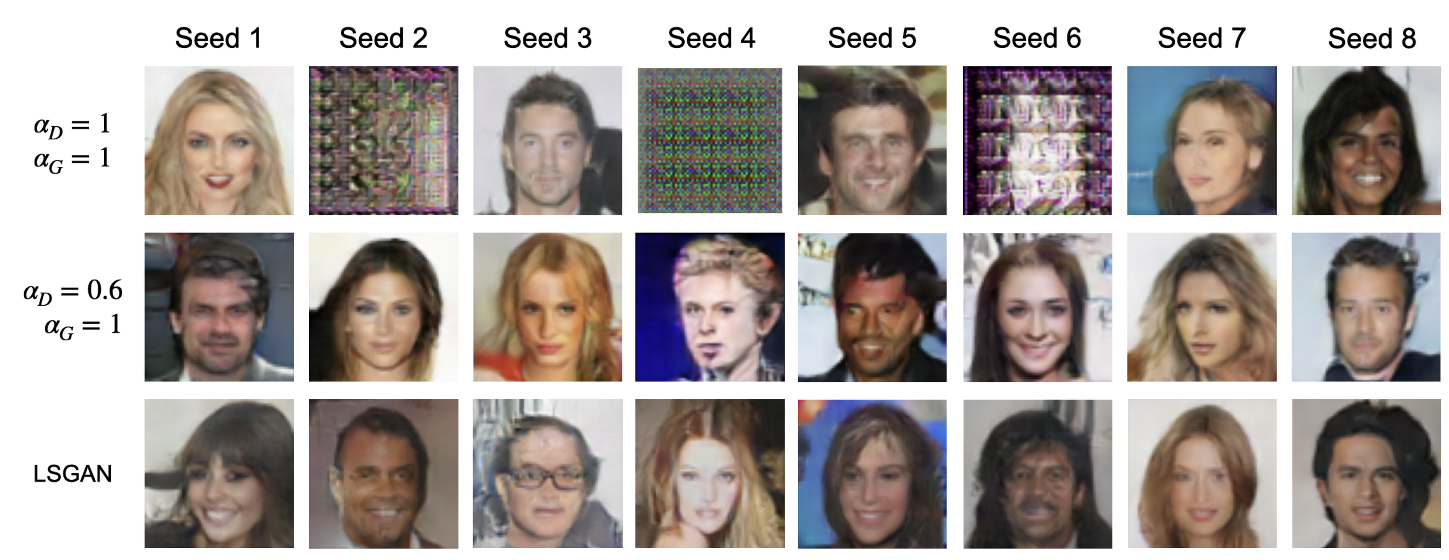


Figure 3: Generated Celeb-A faces from the non-saturating $(1,1)$-GAN (vanilla), the non-saturating $(0.6, 1)$-GAN and LSGAN over 8 seeds when trained for 100 epochs with a learning rate of 5×10^{-4}.

- **LSUN Classroom dataset**: contains over 150,000 classroom images, resized to 112×112

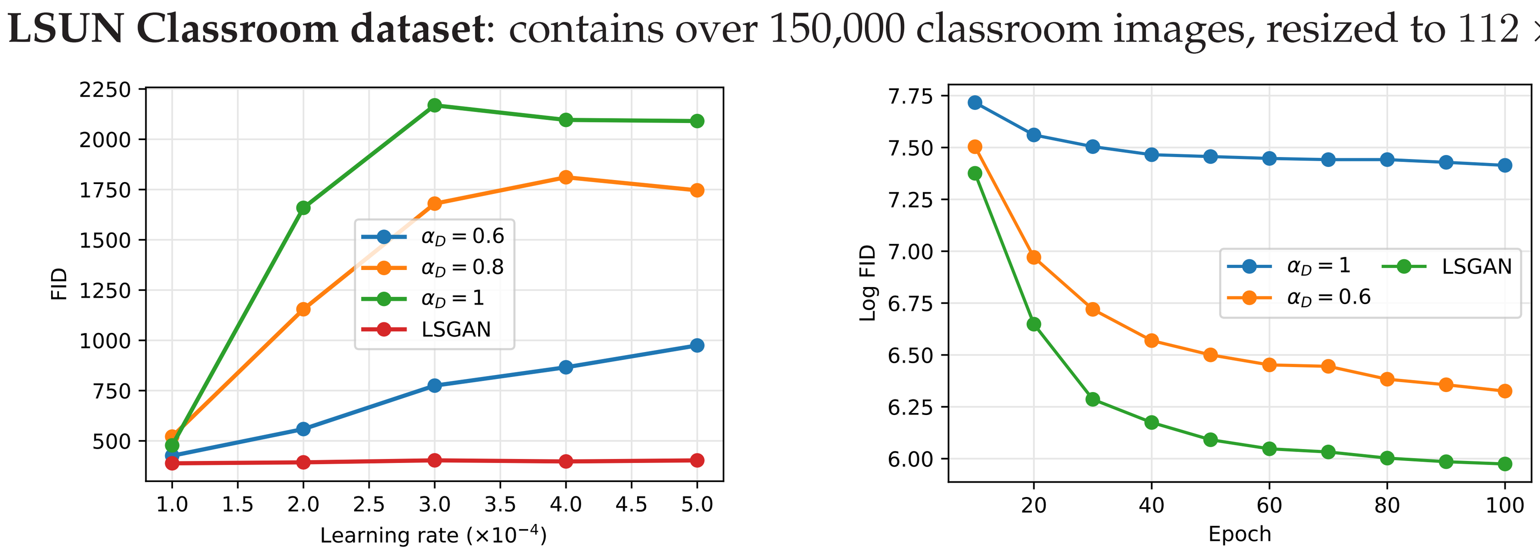


Figure 4: (Left) Plot of FID (smaller is better) averaged over 50 seeds vs. learning rate for a fixed number of epochs (=100) and different non-saturating $(\alpha_D, \alpha_G$=1)-GANs as well as LSGAN. (Right) Log-scale plot of FID over training epochs for the non-saturating $(1,1)$-GAN (vanilla), the non-saturating $(0.6, 1)$-GAN and LSGAN with learning rate 2×10^{-4}.

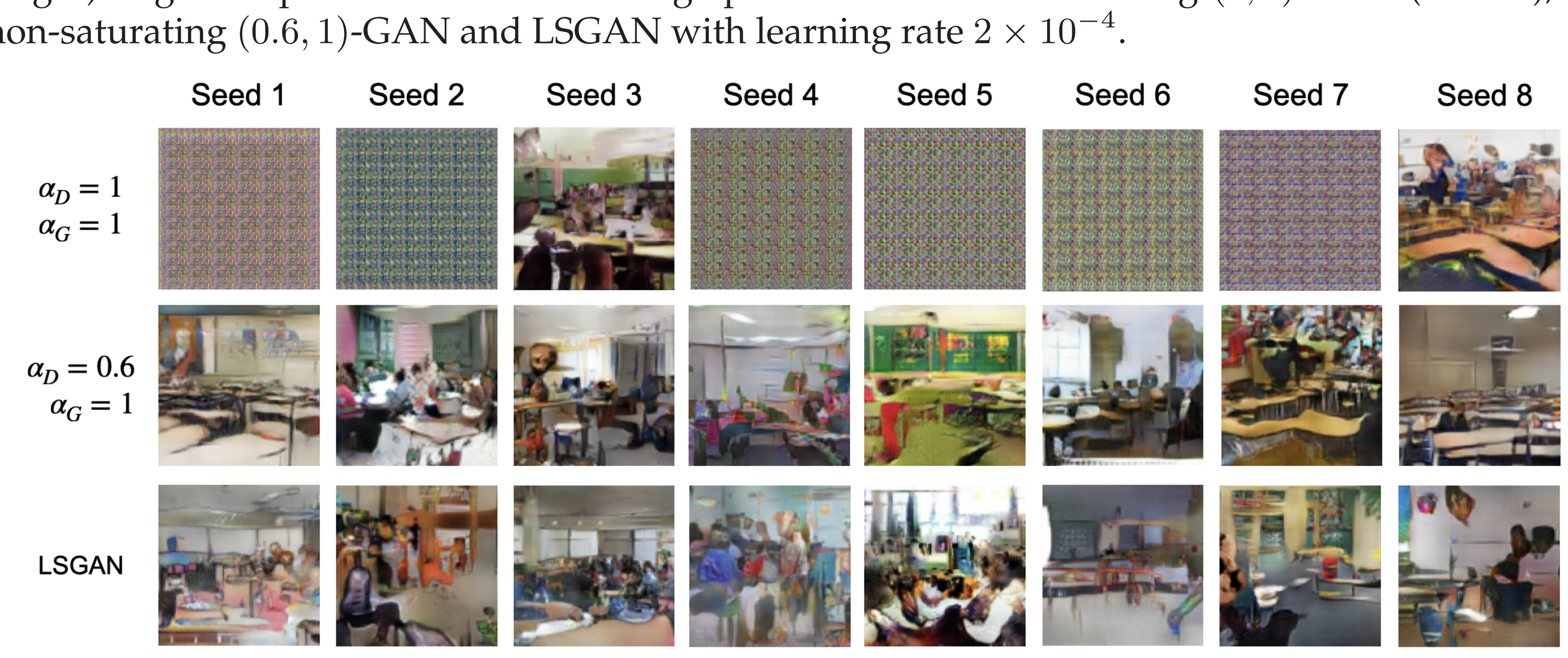


Figure 5: Generated LSUN Classroom images from the non-saturating $(1,1)$-GAN (vanilla), the non-saturating $(0.6, 1)$-GAN and LSGAN over 8 seeds when trained for 100 epochs with a learning rate of 2×10^{-4}.

Takeaway: $\alpha_D < 1, \alpha_G \geq 1$ more robust to hyperparameter initialization, helping to alleviate training instabilities; restricted α_D, α_G ranges make this computationally feasible

Acknowledgements & References

Work supported by NSF grants CIF-1901243, CIF-1815361, CIF-2007688, CIF-2134256, SaTC-2031799, SCH-2205080

[1] Goodfellow et al. Generative adversarial nets. In *NeurIPS*, 2014.
[2] Kurri et al. Realizing GANs via a tunable loss function. In *ITW*, 2021.
[3] Kurri et al. α-GAN: Convergence and estimation guarantees. In *ISIT*, 2022.
[4] Sypherd et al. A tunable loss function for robust classification: Calibration, landscape, and generalization. *IEEE Trans. on Inf. Theory*, 2022.
[5] Welfert et al. (α_D, α_G)-GANs: Addressing GAN training instabilities via dual objectives. In *ISIT*, 2023.
[6] Mao et al. Least squares generative adversarial networks. In *ICCV*, 2017.