Tackling the Generative Learning Trilemma with Denoising Diffusion GANs

Amine Razig Ahmed Khairaldin Zakarya Elmimouni

M2DS IP PARIS

February 24, 2025

Motivation: The Generative Learning Trilemma

- **Key Requirements for Generative Models:**
  - High-quality samples, good mode coverage and fast sampling.

- **The Trilemma:** Mainstream generative models struggle to achieve all three simultaneously.
  - **GANs:** Fast sampling, high quality, but poor mode coverage.
  - **Diffusion Models:** High quality, good mode coverage, but slow sampling.
  - **VAEs:** Good mode coverage, but lower sample quality.

- **Our Focus:** Tackling the trilemma with **DDGAN** (Denoising Diffusion GAN).
  - Combines the strengths of diffusion models and GANs.
  - Achieves fast sampling, high quality, and good mode coverage.

Key Concepts and Formulas of Diffusion models

The Forward Process

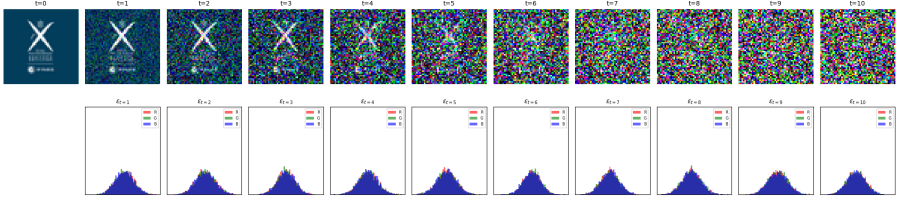
- Gradually add noise to clean images:

$$x_t = \sqrt{1-\beta_t}\, x_{t-1} + \sqrt{\beta_t}\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I)$$

- Transition distribution:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t \mid \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right)$$

- As t increases, x_T becomes nearly indistinguishable from pure noise.



Forward Process: Progressive noise addition (ϵ)

Denoising in DDGANs

Key Idea:

- Reduce the number of denoising steps T by increasing the values of β_t.

Problem:

- When β_t is large, $p(x_{t-1}|x_t)$ is no longer Gaussian.
- The Gaussian assumption does not hold anymore.

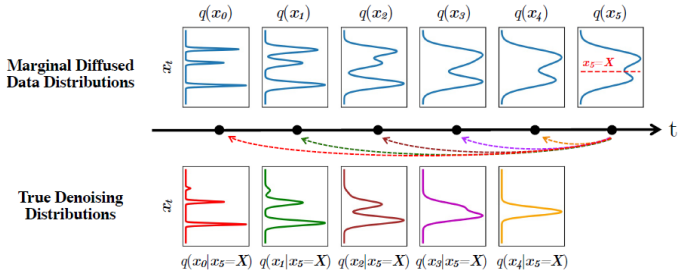


Figure: Effect of increasing β_t on denoising

Conditional GANs

Motivation: Given x_t, we wish to draw a valid x_{t-1} from the distribution $p_\theta(x_{t-1} \mid x_t)$.

A Simple and Direct Approach:

- **Step 1:** Gather many Triplets $\{(x_0, x_{t-1}, x_t)\}$ by applying the forward diffusion process to a large set of images.
- **Step 2:** Train a *conditional GAN* that learns to map x_t (and latent noise) to plausible x_{t-1} samples.

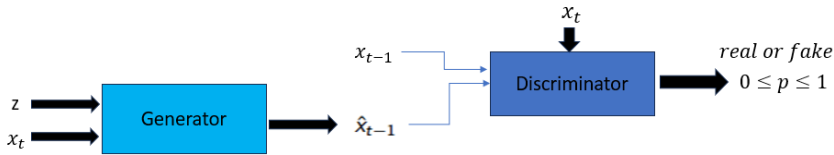


Figure: Condionnal GAN's principle

DDGANs: Bridging DDPM and GANs

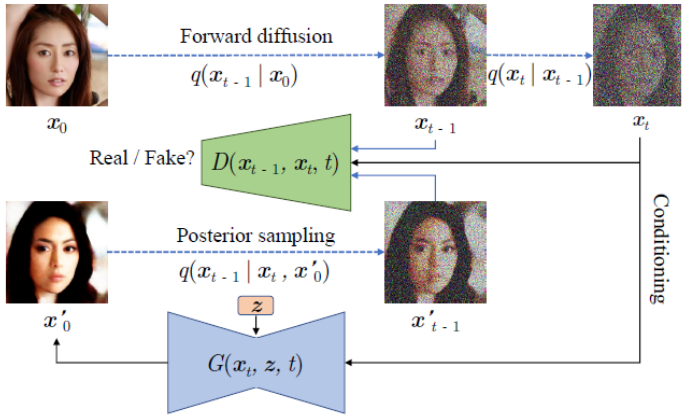


Figure: The training process of denoising diffusion GAN

DDGAN Training Objective

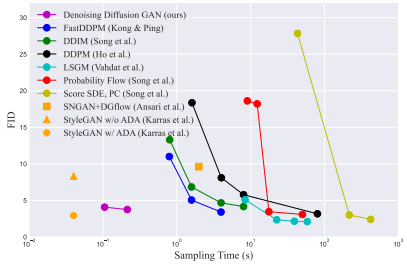
- **Discriminator Training:**
  $\min_\phi \sum_{t\geq 1} \mathbb{E}_{q(x_t)}\Big[\mathbb{E}_{q(x_{t-1}|x_t)}(-\log D_\phi(x_{t-1}, x_t, t)) + \mathbb{E}_{p_\theta(x_{t-1}|x_t)}(-\log(1 - D_\phi(x_{t-1}, x_t, t)))\Big]$.
  - The discriminator's parameter ϕ is updated to associate value near to 1 for real samples and values near to 0 for fake ones.

- **Generator Training (with fixed ϕ):**

$$\max_\theta \sum_{t\geq 1} \mathbb{E}_{q(x_t)}\Big[\mathbb{E}_{p_\theta(x_{t-1}|x_t)}\big[\log D_\phi(x_{t-1}, x_t, t)\big]\Big].$$

  - The generator's parameter θ is updated to produce denoised samples x_{t-1} that the discriminator classifies as real.

Model Comparison on CIFAR-10



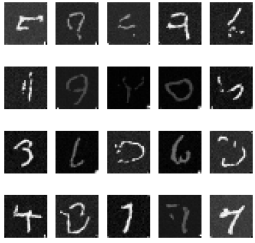
(a) Sample quality vs. sample time trade-off

Model	IS↑	FID↓	Recall↑	NFE ↓	Time (s) ↓
Denoising Diffusion GAN (ours), T=4	9.63	3.75	0.57	4	0.21
DDPM (Ho et al., 2020)	9.46	3.21	0.57	1000	80.5
NCSN (Song & Ermon, 2019)	8.87	25.3	-	1000	107.9
Adversarial DSM (Jolicoeur-Martineau et al., 2021b)	-	6.10	-	1000	-
Likelihood SDE (Song et al., 2021b)	-	2.87	-	-	-
Score SDE (VE) (Song et al., 2021c)	9.89	2.20	0.59	2000	423.2
Score SDE (VP) (Song et al., 2021c)	9.68	2.41	0.59	2000	421.5
Probability Flow (VP) (Song et al., 2021c)	9.83	3.08	0.57	140	50.9
LSGM (Vahdat et al., 2021)	9.87	2.10	0.61	147	44.5
DDIM, T=50 (Song et al., 2021a)	8.78	4.67	0.53	50	4.01
FastDDPM, T=50 (Kong & Ping, 2021)	8.98	3.41	0.56	50	4.01
Recovery EBM (Gao et al., 2021)	8.30	9.58	-	180	-
Improved DDPM (Nichol & Dhariwal, 2021)	-	2.90	-	4000	-
VDM (Kingma et al., 2021)	-	4.00	-	1000	-
UDM (Kim et al., 2021)	10.1	2.33	-	2000	-
D3PMs (Austin et al., 2021)	8.56	7.34	-	1000	-
Gotta Go Fast (Jolicoeur-Martineau et al., 2021a)	-	2.44	-	180	-
DDPM Distillation (Luhman & Luhman, 2021)	8.36	9.36	0.51	1	-
SNGAN (Miyato et al., 2018)	8.22	21.7	0.44	1	-
SNGAN+DGflow (Ansari et al., 2021)	9.35	9.62	0.48	25	1.98
AutoGAN (Gong et al., 2019)	8.60	12.4	0.46	1	-
TransGAN (Jiang et al., 2021)	9.02	9.26	-	1	-
StyleGAN2 w/o ADA (Karras et al., 2020a)	9.18	8.32	0.41	1	0.04
StyleGAN2 w/ ADA (Karras et al., 2020a)	9.83	2.92	0.49	1	0.04
StyleGAN2 w/ Diffaug (Zhao et al., 2020)	9.40	5.79	0.42	1	0.04

(b) Performance of DDGAN vs. diffusion and GAN models

Figure: Paper's experiments on CIFAR-10 dataset

Experiment on MNIST dataset



(a) Generated samples with DDPM

(b) Generated samples with DDG

Figure: Comparison between generated samples DDPM vs. Denoising Diffusion GANs

Conclusion

- Tackling the generative learning trilemma with DDGAN.
- DDGAN uses a complex multimodal distribution to take large denoising steps.
- This model is competitive in the three key points of the trilemma with state-of-the art models.
- Computationally costly model : use of multiple GPUs by the NVIDIA team.

Appendix 1: DDPM Denoising Process

- **Denoising Process:** The true conditional is given by

$$q(x_{t-1} \mid x_t, x_0) = \mathcal{N}\Big(x_{t-1};\, \tilde{\mu}_t(x_t, x_0),\, \tilde{\beta}_t I\Big)$$

where

$$\tilde{\mu}_t(x_t, x_0) = \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\, x_0 + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\, x_t.$$

Since x_0 is unknown during sampling, we estimate it with a neural network $f_\theta(x_t, t)$ (i.e., $\hat{x}_0 = f_\theta(x_t, t)$) and define:

$$p_\theta(x_{t-1} \mid x_t) = \mathcal{N}\Big(x_{t-1};\, \tilde{\mu}_t\big(x_t, f_\theta(x_t, t)\big),\, \sigma_t^2 I\Big).$$

Appendix 2: DDPM and DDGANs Sampling Formulas

- In DDPM, the reverse process is modeled by a Gaussian:

$$p_0(x_{t-1} \mid x_t) = q\big(x_{t-1} \mid x_t, x_0 = f_\theta(x_t, t)\big).$$

  f_θ is modeled by a neural network and q is a normal distribution.

- DDGANs use a conditional GAN to model the denoising distribution.
- The reverse process is defined as:

$$p_\theta(x_{t-1} \mid x_t) = \int p(z)\; q\Big(x_{t-1} \mid x_t,\; x_0 = G_\theta(x_t, z, t)\Big)\, dz$$

- Here, the generator $G_\theta(x_t, z, t)$ predicts an estimate of x_0 based on the current noisy image x_t, the latent variable z (with $z \sim \mathcal{N}(0, I)$), and the time step t.

Appendix: Rewriting the Expectation

- By applying the identity

$$q(x_t, x_{t-1}) \;=\; \int dx_0\, q(x_0)\, q\big(x_{t-1} \mid x_0\big)\, q\big(x_t \mid x_{t-1}\big),$$

we can rewrite the expectation as follows:

$$\mathbb{E}_{q(x_t)\, q(x_{t-1}|x_t)}\big[-\log D_\phi(x_{t-1}, x_t, t)\big] \;=\; \mathbb{E}_{q(x_0)\, q(x_{t-1}|x_0)\, q(x_t|x_{t-1})}\big[-\log D_\phi(x_{t-1}, x_t, t)\big].$$

Appendix 3: Key Concepts and Formulas of Diffusion models

The Denoising Steps: From Noisy x_T to Clean x_0

- Gaussian assumption for $q(x_{t-1}|x_t)$ holds only if:
  - Step size β_t is infinitesimal.
  - Data marginal $q(x_t)$ is Gaussian.

  Reducing denoising steps T breaks the Gaussian assumption.

The Key Idea

Model the denoising process with a **multimodal distribution** using conditional GANs.

Conditional GANs approximate the true denoising distribution $q(x_{t-1}|x_t)$.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan.
Large scale gan training for high fidelity natural image synthesis.
In *International Conference on Learning Representations (ICLR)*, 2019.

[2] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio.
Density estimation using real nvp.
In *International Conference on Learning Representations (ICLR)*, 2017.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio.
Generative adversarial nets.
In *Advances in Neural Information Processing Systems*, volume 27, pages 2672–2680, 2014.

[4] Jonathan Ho, Ajay Jain, and Pieter Abbeel.
Denoising diffusion probabilistic models.
In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

[5] Diederik P Kingma and Max Welling.
Auto-encoding variational bayes.
arXiv preprint arXiv:1312.6114, 2014.

[6] Durk P Kingma and Prafulla Dhariwal