Task 2: Generative Models — VAEs vs GANs

ATML Assignment 1

September 25, 2025

1 Introduction

In this section we will focus on how inductive biases in model purpose or architecture shape how the generative models learn, capture and represent data. The generative models we will discuss are Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAEs optimize an ELBO loss (reconstruction + KL loss) making a structured latent space with smooth interpolations helping in generating reconstructions despite being blurry and lacking detail. GANs on the other hand use adversarial training, where generator tries to generate an output that the discriminator cannot classify as real or fake. This helps GANs produce sharp, realistic samples, however in GANs there's no encoder to structure the latent space making the generations unreliable to some extent.

We will evaluate VAEs and GANs trained on MNIST (28×28), CIFAR-10 (32×32), and CelebA (64×64). Our analysis will include (i) reconstruction quality and how it reflects latent organization. (ii) realism vs diversity tradeoff, (iii) latent-space interpolations and semantnic structure, and (iv) model behavior on OOD inputs.

2 Datasets and Preprocessing

MNIST (28×28 grayscale): Normalized to $[-1, 1]$ with mean=0.5, std=0.5; batch size 128.
CIFAR-10 (32×32 RGB): Normalized channel-wise to $[-1, 1]$ with mean=(0.5,0.5,0.5), std=(0.5,0.5,0.5); batch size 128.
CelebA (178×218 RGB): CenterCrop(178) then resized to 64×64 (no normalization); batch size 128.

3 Models and Training Setup

Model	Optimizer (lr, betas)	Loss
VAE CIFAR-10	Adam (1e-3)	MSE + KL
VAE MNIST	Adam (1e-3)	MSE + KL
VAE CelebA	Adam (1e-3)	BCE + KL (decoder Sigmoid)
DCGAN CIFAR-10	Adam (2e-4, 0.5, 0.999)	BCE-with-logits
DCGAN MNIST	Adam (G:2e-3, D:1e-3)	BCE-with-logits
DCGAN CelebA	–	BCE-with-logits

Table 1: Key hyperparameters and training settings for VAEs and DCGANs.

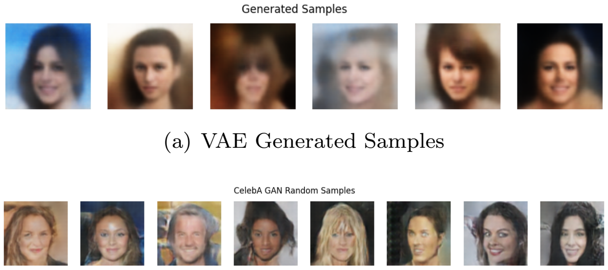
All models use **batch size 128**. **Latent dimensions**: VAE CIFAR-10/MNIST: 64, VAE CelebA: 200, DCGAN all: 64. **Epochs**: VAE CIFAR-10/DCGAN CIFAR-10: 150, VAE MNIST/DCGAN MNIST: 30, VAE CelebA: 20, DCGAN CelebA: pretrained.

4 Results

We organize results by dataset with paired VAE/GAN visuals. All images referenced below are generated by the corresponding notebooks in the repository.

4.1 Visual Quality vs Diversity

GAN samples are sharper images with finer details than VAE. The adversarial loss pushes for realism so textures become more defined as we go through epochs. Also since it was a well pretrained GAN, the mode collapse is avoided and decent diversity is observed, however, as the training samples contained more images of woman, the model becomes slighly biased towards generating images of woman. VAE on the other hand was trained in house, the visuals are smoothed due to the KL term leading to blurry samples. Also since the gender ratio of CelebA is unbalanced towards woman, the model tends to generate more features related to woman.



(a) VAE Generated Samples

(b) GAN Generated Samples

Figure 1: Comparison of generated samples from VAE and GAN on CelebA dataset.

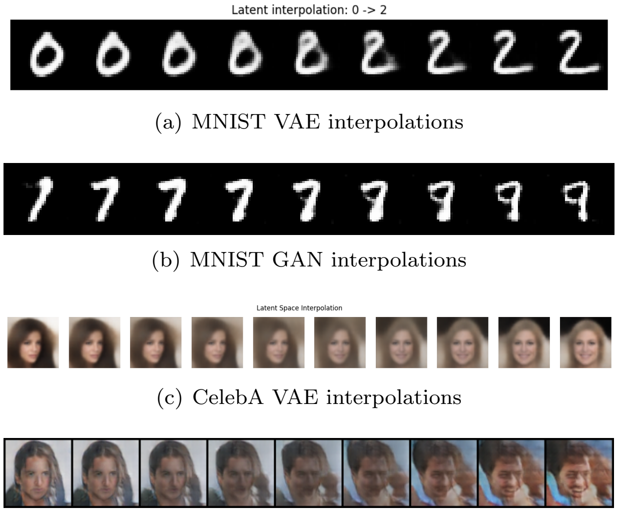
Model	MSE	SSIM	FID	KID
GAN (CelebA)	-	-	97.1	0.015
VAE (CelebA)	0.010	0.677	136.0	0.088

Table 2: CelebA metrics

VAE acheived better reconstrucion performance with lower (MSE = 0.010, SSIM = 0.677) showing that generations are related to input even if blurry, but poor sample quality with higher (FID = 136, KID = 0.088) meaning less realistic samples, despite fine reconstruction. This is because of the bias to cover all modes in data distribution at cost of fidelity. In comparison GAN producing sharper, realistic samples that are closer to data distribution have a lower FID and KID (FID = 97.1, KID = 0.015) consistent with adversarial training bias toward perceptual realism.

4.2 Latent Space Analysis

The VAE MNIST interpolation results of latent space shows a smooth transition from 0 to 2. We can see that the interpolation passes through the digit 8. This also ties with the t-SNE analysis where cluster of 8 (mustard) is in between cluster of 0 (blue) and cluster of 2 (green). This shows that the latent space in VAE is continuous and semantically structured. In contrast, GAN interpolation from 7 to 9 is also smooth with sharp visuals and no discontinuity. GAN interpolation gives higher fidelity samples and avoid the morphing effect (low intensity pixels) as the generator is trained needs to fool the discriminator, the generator generates similar samples from a given input point and around that point as it has learned to fool the discriminator (slight mode bias), when it is given input from a specific subspace of latent space. VAE on the other hand produces lower intensity pixels and shows morphing effect as the latent space is pushed towards being continuous by the KL term. Lower intensity pixels in VAE are shown while traveling through (between) clusters. In the CelebA dataset, the VAE again produces smooth yet blurry interpolation, however in GAN there is irregularity might be due to discontinuity in latent space, though it generates much sharper images due to adversarial loss.



(a) MNIST VAE interpolations

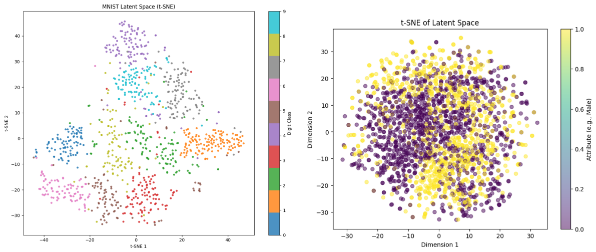
(b) MNIST GAN interpolations

(c) CelebA VAE interpolations

(d) CelebA GAN interpolations

Figure 2: Latent space interpolations for MNIST and CelebA (VAE and GAN).

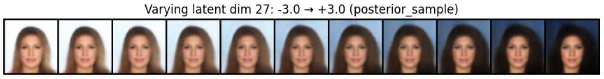
MNIST plot of t-SNE show distinct cluster of number representations in latent space where clusters like 1 and 7 lie closer to each other due to similarity in how they are written, and clusters for 0 and 1 appear far from each other as they are dissimilar in writing, meaning that the latent space is continuous and semantically structured. The t-SNE for VAE latent trained on CelebA show the woman distributed on the entire normal while the men mainly clustered around a point. This means that model encoded greater variability in female faces (consistent with the dataset bias of gender ratio), while it learned less variability in men features as model saw fewer examples of men.



(a) MNIST VAE t-SNE

(b) CelebA VAE t-SNE

Figure 3: MNIST VAE latent space



(a) Latent dimension sweep with auto-dim selection and flexible anchor

Figure 4: Latent traversal on less important dimensions

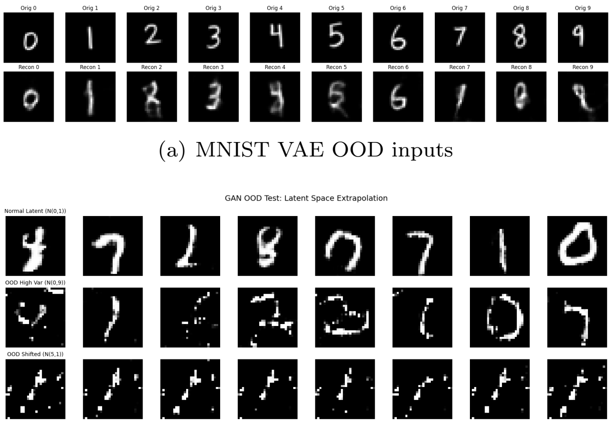
The figure above shows latent sweep while keeping the most responsive dimension constant. We can observe that the background and hair colour change as we traverse while the important dimension of facial features are the same. This signifies good learning across different dimensions as model learned important features and less important features distinctly. An important insight is the travsal varies the skin colour which indicates that features such as eyes, nose, face size and rest of facial features are represented distintly in latent space than skin colour of the person.

4.3 OOD Robustness Analysis

We tested OOD robustness of the VAE trained on MNIST by reconstructing hand written images through the model. VAE reconstructed good reconstructions with surprisingly less average reconstrucion error (23.35) than IID test data (30.36). The reason for this can be that the digits are really well written and the VAE is well trained and latent space is semantically organized well as can be seen by the t-SNE. The OOD inputs might have landed close to regions from where decoder produce plausible samples. Also the reconstrucion loss for simple digits (like 1 with a single stroke) is much lesser and reconstructions are better than more complex digits like 2 or 8. The blur on the OOD reconstrucions is higher as VAE's KL-regularized decoder tends to average inputs toward a smooth latent representation.

In case of GAN, we tested the model on high variance N(0, 10) and shifted N(5, 1) inputs. The outputs for high variance are somewhat facelike but very distorted, indicating that the generator struggles with latent magnitude far beyond N(0, 1). For the shifted inputs the generator produces pure noise, without any meaning indicating that the generator cannot generalize on out of trained latent region. This means that GAN are not robust to extrapolation as generator is tightly bound to the standard normal. This tells about the bias toward high fidelity constructions within the training bounds but poor generalization specially outside the training support.

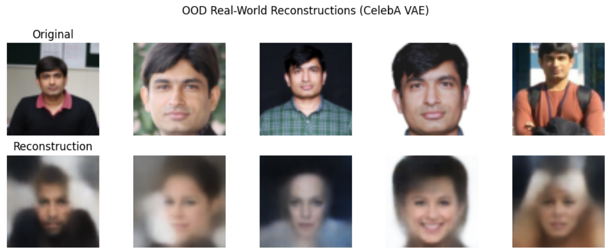
We also tested OOD robustness of the VAE trained on CelebA dataset by providing real world images. The VAE produced outputs with blurred background and edges, and average facelike structure unlike the real image with high average reconstrucion loss (209.2) indicating anomally detection. Five images of a man were given as input where four were distinct poses and one was a cropped version. Though all input images were of the same man, VAE reconstructed four out of 5 female faces and just one male face. Moreover every reconstructed face was different from other indicating reconstructions did not preserve identity and tended to 'hallucinate' plausible but mismatched samples rather than faithfully reproducing the true input. This reflects two inductive biases: i) the KL prior pushes reconstructions towards dominant regions of latent space, ii) the CelebA dataset unbalanced gender ratio biases the model to generate female-like ouputs. As a result, the VAE generalizes poorly to real-world OOD inputs and exhibits mode averaging rather than identity preservation.



(a) MNIST VAE OOD inputs

(b) MNIST GAN OOD outputs

Figure 5: OOD robustness on MNIST: VAE (up) and GAN (down)



(a) CelebA VAE OOD inputs

Figure 6: CelebA VAE OOD Results

5 Training and Stability

The VAE training on MNIST dataset was relatively smoother, than the training on CelebA dataset or training of GAN. First few epochs tuned KL-weight while loss reduced steadily. VAE trained on CelebA produced many challenges due to unbalanced gender ratio of dataset, training had to be just sufficient that model learns important features while not biasing toward generating only female faces. The training was run for reduced epochs (20) and monitored to get the best result. The KL-weight had o be kept optimal as high KL-weight produced overly blurred outputs and low KL-weight made the latent space collapse, and interpolations losing smoothness.

GANs being more delicate were harder to train due to mode collapse. When we were training GAN on CIFAR-10, the discriminator loss falling to zero indicating mode collapse while training was avoided later by decreasing learning rate. The adversarial setup led to oscillating generator and discriminator losses, and tuning was required to avoid either player overpowering the other. This reflects the inductive bias vs. flexibility trade-off: GANs achieve sharper, more realistic samples by focusing on distributional matching rather than explicit density estimation, but this makes training more unstable and prone to collapse compared to VAEs. This reflects the inductive bias vs. flexibility trade-off: GANs achieve sharper, more realistic samples by focusing on distributional matching rather than explicit density estimation, but this makes training more unstable and prone to collapse compared to VAEs.

6 Conclusion

In conclusion, our experiments on these generative models highlight distinct inductive biases in VAEs and GANs. Where VAEs are easy to train and learn entire true distribution, they produce blurry outputs. GANs while producing sharper, high quality images suffer with mode collapse while training and can lack diversity. Together, these findings illustrate the central fidelity-diversity trade-off. Understanding these biases not only explains the practical strengths and weaknesses of each model, but also informs the design of future generative approaches that seek to combine stability, representation quality, and realism.