Towards Controllable High-Quality Image Generation

Team-3

Rajiv Chitale[1] Vedant Bhandare[2] Umanshiva Ladva[3] Nitya Bhamidipaty[4]

[1]CS21BTECH11051
[2]CS21BTECH11007
[3]AI22BTECH11016
[4]CS21BTECH11041

Introduction

- **Generative models** have shown remarkable progress in synthesizing high-quality images across domains such as faces, medical scans, and art.
- Despite this progress, real-world applications like:
  - **Inpainting** (filling missing regions),
  - **Object removal**, and
  - **Attribute-based generation**

  demand a high degree of **controllability** over generated content.
- **Key bottlenecks** include:
  - Entangled latent spaces with no semantic separation
  - Lack of modular control (e.g., pose vs expression)
  - Trade-offs between quality, speed, and flexibility

Motivation

- **Current Limitations**:
  - **VAEs**: Struggle with blurry outputs due to pixel-level losses.
  - **GANs**: High quality, but lack interpretable or structured latent spaces.
  - **DDPMs**: SOTA image quality, but expensive and hard to control.
- **Our Motivation**:
  - Combine the interpretability of VAEs with the quality of DDPMs.
  - Improve **semantic disentanglement** for user-guided generation.
  - Reduce **inference and training cost** via partial denoising.
- **Goal**:

Enable *fine-grained control* over specific image attributes (e.g., gender, hair, pose), while maintaining high perceptual quality.

Literature Review: Foundational Models

- **Variational Autoencoders (VAEs)**
  - Probabilistic latent variable models with efficient training.
  - Enables latent space interpolation and semantic structure.
  - *Limitation:* Often generate blurry outputs due to pixel-wise losses.
- **Generative Adversarial Networks (GANs)**
  - Adversarial training between generator and discriminator.
  - Capable of generating highly realistic images.
  - *Limitation:* Poor latent controllability and training instability.
- **Denoising Diffusion Probabilistic Models (DDPMs)**
  - Models data distribution through a gradual noising-denoising process.
  - Achieves SOTA image quality.
  - *Limitation:* Expensive to train and slow during inference.

Literature Review: Improving Control

- **DiffuseVAE (Kushagra Pandey et al)**:
  - Combines VAE and DDPM for better sample quality and training efficiency.
  - Preserves VAE-style interpretability while reducing DDPM's cost.
  - *Baseline for our ContraVAE + partial denoising strategy.*
- **Conditional GANs (Mehdi and Simon et al)**:
  - Enable attribute-based generation via label embedding or auxiliary inputs.
  - Improve interpretability and controllability over features like class, color, pose.
  - *Laid groundwork for our conditional contrastive setup.*
- **ContraGAN (Minguk and Jaesikk et al)**:
  - Introduces a contrastive loss on discriminator features to improve semantic separation.
  - Enhances controllability by maximizing inter-class distance in latent space.
  - *Motivated our use of contrastive loss in VAEs.*

Literature Review: Improving Control

- **Latent Diffusion Models (LDM)(Robin Rombach et al)**:
  - Propose diffusion in a compressed latent space instead of pixel space.
  - Significantly reduces computational load while preserving image quality.
  - *Inspired our partial denoising in VAE latent space.*
- **ControlNet(Lvmin Zhang et al)**:
  - Text-to-image models can't handle spatial cues - *edges*, *poses*, *segments*.
  - ControlNet runs a trainable copy alongside frozen pretrained SD.
  - Control input c feeds into the trainable copy.
  - c is reduced to SD's latent space - keeps computation low.

Datasets Used

- **MNIST**:
  - Used for initial validation of model controllability and reconstruction quality.
- **CelebA (Liu, Ziwei et al)**:
  - Large-scale face attributes dataset with over 200K celebrity images.
  - We used a subset of 80K images to balance training efficiency and quality.
  - Rich in facial attribute variability — useful for testing attribute-level control.
- **dSprites (TensorFlow Datasets)**:
  - Procedurally generated 2D shapes dataset with 6 independent ground truth factors: *shape*, *scale*, *rotation*, *x-position*, *y-position*, *color*.
  - Ideal for latent space disentanglement analysis.
  - Allows controlled manipulation of generative factors during evaluation.

Feature Extraction from latent space

- Trained VAE on two datasets - celebA and MNIST.
- Attempted to extract features from learnt latent space using **Principal Component Analysis (PCA)**
- **celebA**
  - Background Removal
  - Hair Color
  - Skin Tone
- **MNIST**
  - Thin & Bold font strokes

Feature Control using PCA

- **celebA**

Figure: Background Removal

Figure: Changing Hair Color

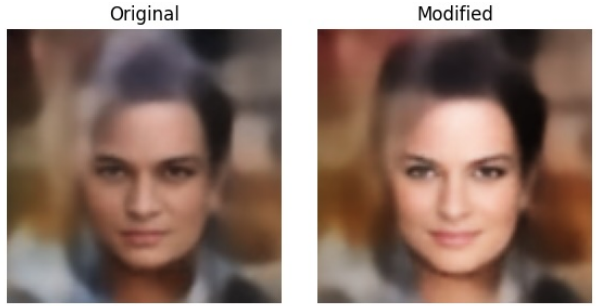

Figure: Changing Skin Tone

Feature Control using PCA - continued

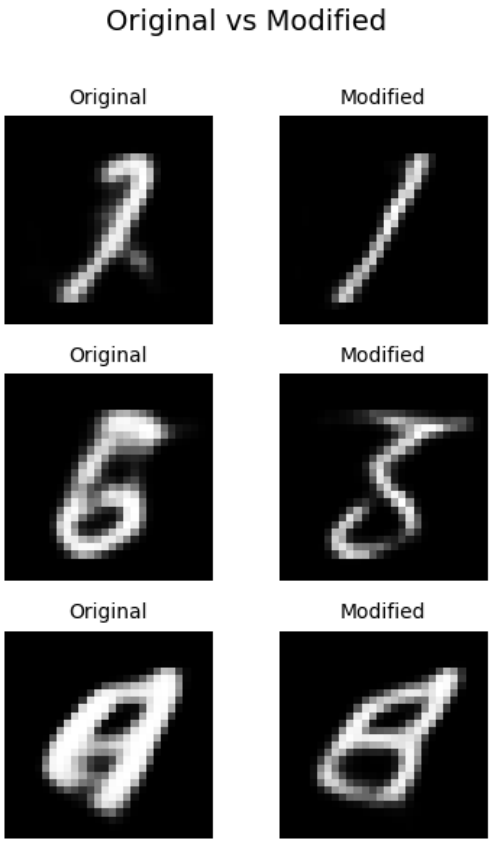

Figure: Reducing Stroke Thickness

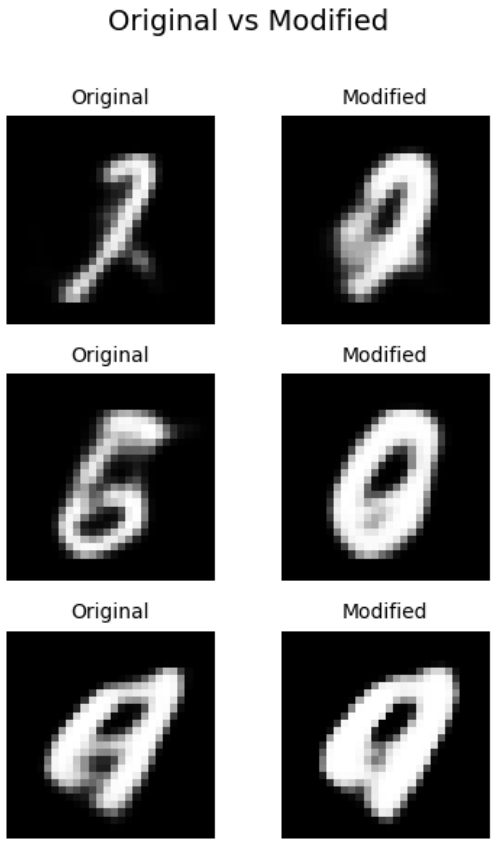

Figure: Increasing Stroke Thickness

Feature Control using PCA - Observations

- PCA-based edits occasionally produced the intended changes in stroke thickness.
- However, these manipulations often altered the overall shape of the digit, sometimes turning it into a different digit.
- This suggests that principal components capture entangled factors, not isolated features.
- VAEs do not enforce disentangled representations in the latent space.
- *We switch to* **contrastive learning** *- semantically meaningful and disentangled latent representations for effective control.*

Conditional GAN on MNIST

- **What is cGAN?**
  - Conditional GANs **(Mehdi Mirza et al)** extend GANs by incorporating label information into both the generator and discriminator.
  - Enables class-conditioned image generation, improving control over output categories.
- **Our Experiment**:
  - Trained a cGAN on the MNIST dataset.
  - Digit labels (0–9) used as conditional input to both generator and discriminator.
  - Generated class-specific digit images with improved clarity and diversity.
- **Key Takeaway**:
  - Conditioning improves controllability without sacrificing generation quality.

CGAN

Figure: Outputs after conditioning on labels

ContraVAE: Contrastive Variational Autoencoder

- **Motivation**:
  - GANs (incl. ContraGAN lack the ability to encode images into latent representations.
  - Standard VAEs suffer from entangled latent spaces—limiting interpretability and control.
- **Our Approach: ContraVAE**
  - Combines a VAE with contrastive loss to improve latent space disentanglement.
  - Two augmented views of an image are encoded; a contrastive loss pulls their latent vectors closer.
  - Total loss = VAE loss (reconstruction + KL) + $\lambda \times$ contrastive loss.
- **Why Contrastive Learning?**
  - Promotes separation of latent factors (e.g., shape, scale).
  - Leads to a more structured and interpretable latent space.

ContraVAE

Figure: Shape control

Figure: Position control

Figure: Entangled position and shape

VAE-Diffusion Hybrids

- Use VAE for structure, jumping to timestep T' of diffusion
- Denoise final T' timesteps for sharp details
- Two variants
  - Noisy VAE + Partial Diffusion
  - VAE + Noise + Partial Diffusion

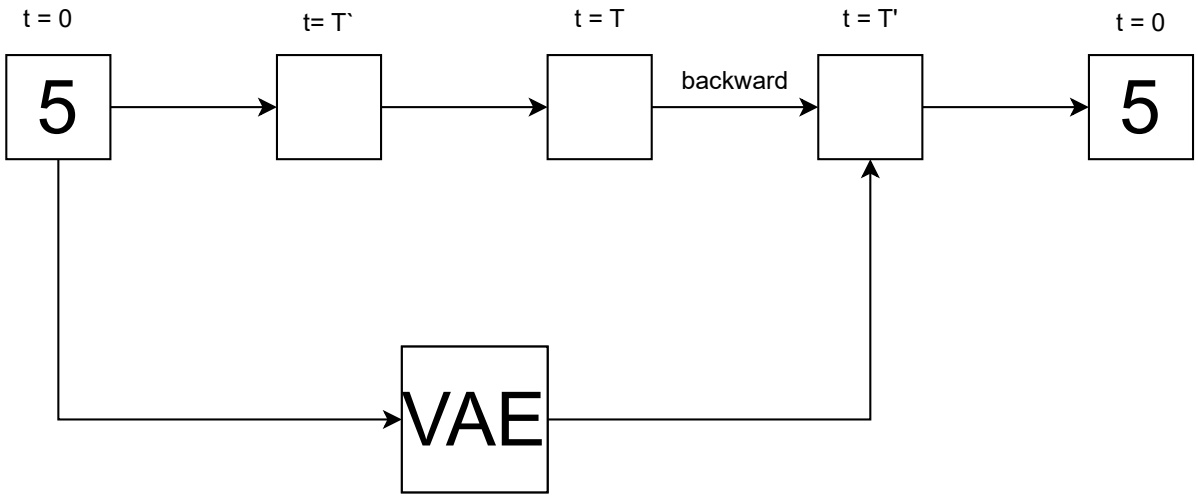

Figure: Noisy VAE + Partial Diffusion

Outputs from Partial Denoising

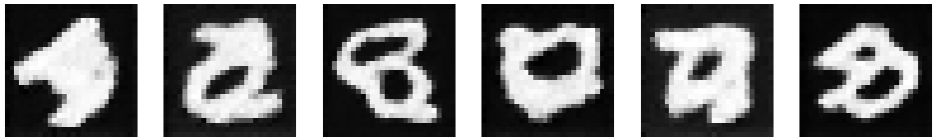

Figure: DDPM

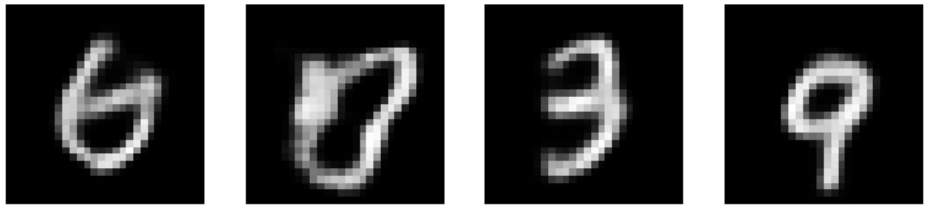

Figure: Noisy VAE output

Figure: Outputs after Partial Denoising

Investigating VAE for Frequency Domain

- Discrete Fourier Transform produces frequency spectrum
- Can this representation allow a VAE to learn sharp details?

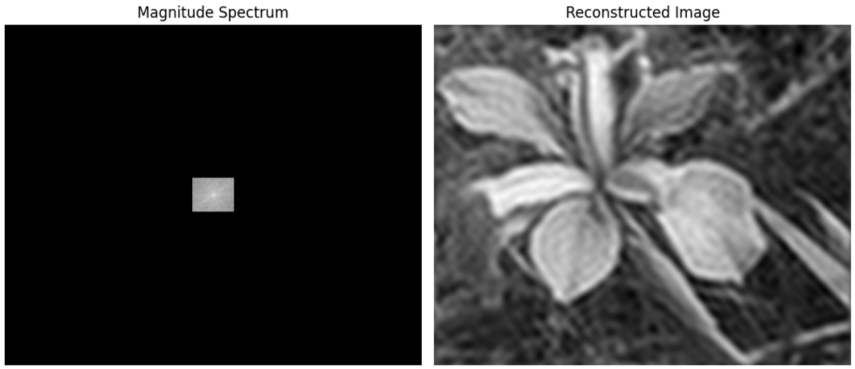

Figure: Low Frequencies - overall structure

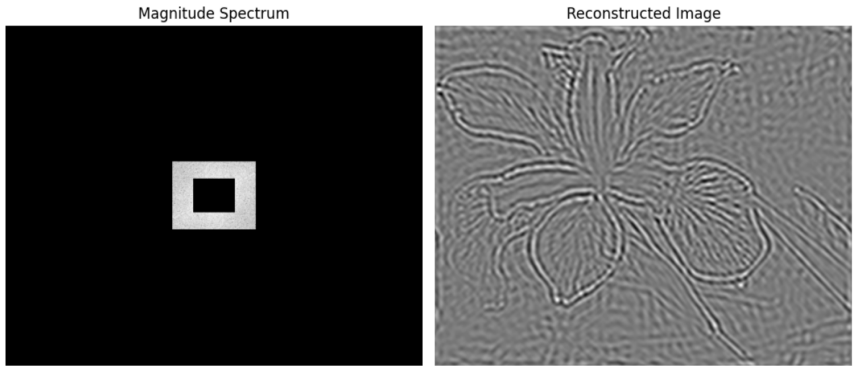

Figure: Medium Frequencies - sharp details

Issues with VAE for Frequency Domain

- Significantly affected by noise
- Operations like 3x3 blur or shift distorted image
- Trained VAE was able to reproduce blurry images
- But generated images were not realistic

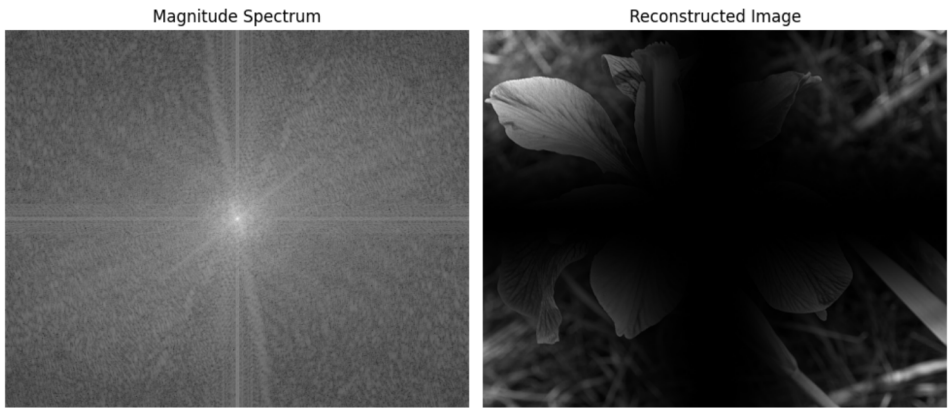

Figure: Inverse DFT after 3x3 blur

Key Learnings

- **Technical Insights**:
  - Latent space structure directly affects controllability
  - Contrastive Learning improves interpolation and disentaglement
  - Partial denoising reduces inference steps, can be used for style
  - Frequency domain is very sensitive to errors, not suitable for models
- **Collaborative Skills**:
  - Synchronous training on distributed systems
  - Debugging instability (NaN gradients, mode collapse)

Team Contributions

- **Rajiv**:
  - **Survey:** VAE, DDPM, DiffuseVAE
  - **Experiments:** VAE + Partial Diffusion, Frequency Domain VAE
- **Vedant**:
  - **Survey:** Latent Diffusion Models, Control Net
  - **Experiments:** Latent space evaluation and visualization
- **Umanshiva**:
  - **Survey:** GAN, Conditional GAN, ContraGAN
  - **Experiments:** Contrastive learning integration(for VAE) and conditional GANs
- **Nitya**:
  - **Survey:** VAE-GAN, StyleGAN
  - **Experiments:** Implementing DiffuseVAE and Feature based contrastive (for unconditional) GAN