Recognition and Interpretation in Dysfunctional Image Generation

Zihan Zhou[1] and Rui Guo[1]

Department of Computing Science, University of Alberta, Edmonton, AB, Canada
{zhou9,rg5}@ualberta.ca

Abstract. Recent years saw many key breakthroughs made in image generation domains with great potential witnessed by deep generative approaches. However, systematic research for the potential "dysfunctions" is still left blank. The research project is designed to focus on a few potential topics of dysfunctional issues such as "latent feature loss", "unbalancing data generation", and "vanishing divergence". This project explores the marginal fields of image generation techniques and provides insights into further advancement of the existing deep generative processes.

Keywords: Image Generation · Dysfunctional Issues · Latent Space.

1 Introduction

The cumulative maturation of pre-training models, multi-modal techniques, deep generative algorithms and other related applications significantly contribute to AI-based designing and creating development (see Figure 1) and lead to the explosion of *Artificial Intelligence Generated Content* (AIGC). However, there are

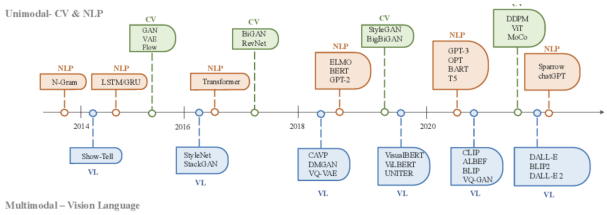

Fig. 1: The technique development about AIGC [3].

still many dysfunctional-related unknowns and issues left. For example, Modern generative algorithms highly rely on the innovation of neural networks, even though researchers can not easily describe how the process works and why "neural networks can be simply applied to solve all kinds of image generation tasks". There is scant detail provided on how the generative model trains to produce

flawed images; the role of individual feature variables within the intricate learning environment remains unclear; and the nuances of what is gained and what is sacrificed in the "encoding-decoding" sequence are not thoroughly explored.

This research project shifts the focus outer the margin of the image generation field to reveal how they function in detail, and dig deeper into the inner space of the deep generative models to discover why dysfunction issues occur. We analyzed the dysfunctional cases from both experiments and theories, from the “implicit latent space" and the “explicit data generations". We examine the “marginal" and “central" dysfunctional cases by sampling generated data points latent space of *Variational autoencoder* depending on two considerations from both “human cognition" and “quantitative metrics" and explain these dysfunctional phenomena from three degrees. This paper provides insights into the dysfunctional interpretation and recognition, and meanwhile provides probable suggestions for further enhancement and fixing of the deep generative models.

2 Existing Works

According to the literature review, many existing researches are done related to the optimization and utilization of image generation models. Some brief overviews of these researches are provided below that focus on two principal topics: the first is some methods of *Latent Space Exploration* and then some enhancements about *Image Generation Estimations*. These existing studies provide notable insights for our study in dysfunction recognition and interpretation.

2.1 About Latent Space Exploration

In 2017, an invention based on the exploration of the latent space of the Generative Adversarial Network was introduced by Bojanowski et al in 2017 in their paper named “Optimizing the Latent Space of Generative Networks" [2]. The authors offer a new framework for training deep convolutional generators depending on simple reconstruction losses called *Generative Latent Optimization* (GLO). The new technique processes the optimization of the generative networks of GANs much differently from the mainstream strategy. The experiment of GLO on GANs provides noteworthy insights into the exploration of “why GANs can do so well" in image generation tasks: this suggests that the success of GANs may not entirely depend on the powerful adversarial nature of their optimization problem and opens up new directions for researching in the domain of generative networks.

The year 2017 also saw some exploration of the oddity of the latent space. Arvanitidis, Hansen, and Hauberg conclude existing works and provide insights about some potential enhancement based on the curvature of deep learning models in their paper *Latent Space Oddity: on the Curvature of Deep Generative Models* [1]. They explained that the nonlinearity of the generator network can lead to a distorted situation in the input variable space, which can be represented by the stochastic Riemannian metric under mild conditions. They proved that the

distances and interpolants are significantly improved under this metric, which in turn brings better performance of the probability distributions in the latent space, such as sampling, and clustering. Additionally, they also emphasized the under-expectation issue of variance estimation led by current generators, they constructed a novel generator architecture for more precise variance estimates. All of their works in this paper highlight the importance of considering the relationship between model oddity and generative performance, which provides a new benchmark for generative model analysis.

Sudipto et al's research [13] focused on the clustering usage of the Generative Adversarial Networks. They argued that the latent space of the Generative Adversarial Network cannot maintain the cluster structure. To explore this and provide insights about the unsupervised learning application, they introduce a method called *Cluster Generative Adversarial Network* (ClusterGAN), which can make efforts to sample latent variables from a mixture of one-hot encoded variables in the continuous latent space. The new approach helps facilitate the clustering process in the latent space by mapping to an inverse network structure. The research outcomes indicate that the latent space interpolation can be saved across classes, even if no discriminator is exposed to these vector spaces. Mukherjee et al's work highlights the potential of Generative Adversarial Networks' extension application in clustering, presenting outstanding performance on both the synthetic dataset and the real dataset compared to other existing clustering baseline methods.

More research on the optimization of the latent space was done by Hu et al in 2023 [8]. They provide new insights into the process of modelling the latent space of the deep generative model. They initially introduce a strategy of minimizing the "distance" between the latent distribution and the data distributions. The optimization process could lead to a reduction in the generator's computational complexity. Besides, a new two-stage training strategy named *Decoupled Autoencoder* (DAE) was also raised in this paper: the dedication contributes to optimizing the latent distribution in order to improve the performance of the generative model. The examinations based on multiple models like diffusion transformers and *vector Quantized Generative Adversarial Networks* (VQGAN) all result in positive results with significant improvements in sample quality and a fall in complexity.

The exploitation of latent space diffusion models on wider application domains results in a unique image restoration technique called *Refusion*, which was raised by Ziwei et al (2023) [12]. The authors make the diffusion model more robust by adjusting several aspects such as network architecture, noise level, denoising stepsize, training strategies and the optimizer design. The careful tuning of these hyperparameters leads to better performance on both distortion and perceptual scores. Additionally, another breakthrough raised is to have a proposal of a U-Net-based latent space model in order to perform a diffusion process based on low-resolution information from the original decoding inputs. The paper argues that their U-Net-based compression approach presents more stable and appropriate recovering accuracy compared to the existing most

common method of *Variational Autoencoder - Generative Adversarial Network* (VAE-GAN). By simply changing the datasets and slightly altering the noise network, the latent model of Refusion can handle large-size images and assemble satisfactory results on types of restoration tasks.

Some recent studies about the latent space in large language models (LLMs) provide a theoretical exploration of the emergent ability and some connection to its corresponding latent semantic representation [9]. The researchers found a sparse joint distribution in the latent space that is heavily peaked according to the strong correlation between semantics and their underlying meanings, and these peak values match the marginal distribution of speeches due to the sparsity. Jiang differentiates languages as either unambiguous or epsilon-ambiguous, to see how quantitative results are presented in demonstrating the emergent abilities of the generative models. The exploration of the sparse joint distribution of languages and its latent position in LLMs contributes to the ongoing research in the field of computational linguistics and deep generative models.

2.2 About Image Generation Estimations

The quality of the generated images can be reasonably assessed both qualitatively and quantitatively. Quantitative assessment methods include Inception score, Fréchet Inception Distance, Mode Score, and other methods. For the generated images, we mainly consider two factors: the clarity of the images and the diversity of the images. The lack of picture clarity is generally due to the lack of expressive ability of the network, which requires the use of a better or more complex network structure, while the lack of picture diversity is most likely due to the selection of the loss function or the training method, the common ones are mode collapsing and mode dropping. Mode Collapsing means that some duplicate results often appear in the generated images. Mode Dropping means that some modes are missing, which also leads to a lack of diversity.

A popular metric is the Inception Score (Salimans et al., 2016)[14], which uses an external model, the Google Inception Network, to assess the quality and diversity of generated images. However, in addition to the clarity and diversity of GAN-generated images reflected by IS, many problems are worth considering. On the one hand, the Inception Score is very sensitive to the internal weights of the neural network. Pre-trained networks with different frameworks achieve the same classification accuracy, but due to slight differences in their internal weights, the Inception Score can vary greatly. On the other hand, values are more disturbed by sample selection and are not suitable for use on datasets with large internal variations.

The nature of the problem with the Inception Score is that only the generated sample is considered in the calculation of the IS, not the real data, i.e., the IS does not reflect the distance between the real data and the sample. That is, a better evaluation of the generative network requires the use of a more efficient method to calculate the distance between the true distribution and the generated samples. So the basic idea of Fréchet Inception Distance[6](FID) is to use the convolutional feature layer of the Inception network as a feature function

and use the feature function to distribute the real data and the generated data distribution modelled as two multivariate Gaussian random variables. It first uses the Inception network to extract features and then uses a Gaussian model to model the feature space and then goes on to solve for the distance between two features; a lower FID means a higher quality and diversity of images. Compared to IS, FID is more robust to noise, but it still does not solve the problem of overfitting on large-scale datasets, and feature extraction-based methods can only be evaluated based on the presence or absence of features, not the relative spatial location of features.

Mode Score[4] is an improved version of the Inception Score, with the addition of a measure of the similarity of the probability distributions of the predictions of the generated and real samples. Kernel Maximum Mean Discrepancy[5] (Kernel MMD) first chooses a kernel function k which maps the sample to Reproducing Kernel Hilbert Space (RKHS). The smaller the MMD value, the closer the two distributions are. It can be a certain measure of the superiority of the image generated by the model with a small computational cost.

1-Nearest Neighbor Two-Sample Test[11] was introduced by Lopez-Paz and Oquab in 2016. The 1-Nearest Neighbor classifier is almost the perfect metric for evaluating GANs. Not only does it have all the advantages of the other metrics, but its output score is in the interval [0, 1], similar to the accuracy in classification problems. This metric obtains a perfect score when the generated distribution matches the true distribution perfectly. It is characterized by the fact that if the feature space is chosen appropriately, it can be effective.

3 Methodology

In this section, we explore how to identify the “dysfunctional" issues that occur in image generation from three views: marginal performance, in-cluster performance, and mathematical proof. Beyond the generation patterns understanding and empiricism-based analysis, some generation quality validation benchmarks are involved. We seek to model the latent space inside deep generative models through clustering methods and evaluate their performance by computing the FID value of the corresponding images.

3.1 Latent Space Exploration

Latent space exploration is a critical research direction related to the deep generative models. Researchers attempt to understand how the latent space is constructed by the “encoder phase" of the generation model and how each generated data point in the latent space is mapped by the “decoder phase" to form a novel image generation. This section introduces the methods to extract, model and analyze the latent space of VAEs.

Generation Patterns Understanding Before digging into the in-depth reasons for poor quality images produced by the mature image generation models,

it is mandatory to first recognize the generation designs and explain why we can create new data points from some provided image data set. As we reviewed in section section 2, there are many VAE variants, and all these models are developed from the identical origin called *Autoencoder*.

AE *Autoencoder* (AE) introduced by Hinton and Salakhutdinov [7] is a neural network that learns the identity function using an unsupervised approach: the data is first efficiently compressed, and then the original input is reconstructed. It consists of an encoder and a decoder. However, Autoencoder still has some drawbacks, such as layer-by-layer training that makes the model training time longer and the fact that Autoencoder is an unsupervised learning that does not describe the features it learns well in a physical sense.

VAE A new model called *Variational Autoencoder* (VAE) was presented by Kingma and Welling[10] in 2013. Unlike Autoencoder, given an input sample x, VAE expects a latent distribution, not a fixed latent representation. The main contribution of Kingma's work is to combine variational inference and auto-decoders, which extends the autoencoder model from an unsupervised data degradation and reconstruction model to a generative model by introducing hidden variables During training, the model minimizes both the *reconstruction error* and the *KL scatter* of the latent variables (p.s., more discussions refer to subsection 4.4), which allows the model to learn the distribution of the data. This model provides new ideas and methods for research in deep learning and generative modelling.

Difference from the classic image processing model of AE, the "variational" design in latent space allows model VAE to process efficient sampling and generation of new data points. The "variational" aspect of VAEs came from the approximation method of the *variational inference*, which involves approximating the posterior distribution of the latent variables using a learned distribution in the latent space. This measure helps to process and manage the marginal areas between two distributions. In simple terms, AE can decode only the "valid" data points inside the latent space and treat all others as "invalid" cases, but VAE learned to handle the "invalid" marginal areas around the valid data points and reasonably generate data from the "invalid" nodes rely on mixed information from the "valid" points nearby. In this way, sampling the special data points from the marginal area may help explain what happened there and what we can do to better exploit or get rid of the influence of the marginal conditions.

ELBO The generation model of VAE (p.s., the similar idea for other image generation models) can be expressed in a log likelihood expression of

$$\ell(\theta; D) = \sum_{i=1}^{N} \ln p_\theta(z^{(i)}, x^{(i)}) \tag{1}$$

where θ indicates the parameters, $z^{(i)} \in Z$ presents the latent variables of the model, and $x^{(i)} \in X$ presents the input training information. The dataset $D = \left\{(z^{(1)}, x^{(1)}),\ \cdots, (z^{(N)}, x^{(N)})\right\}$ should collect all these information, but actually

the latent variables in Z are unknown. In this way, we have the following equation

$$\ell(\theta; D) = \sum_{i=1}^{N} \ln p_\theta(x^{(i)}) = \sum_{i=1}^{N} \ln \int_z p_\theta(x^{(i)}, z) dz \tag{2}$$

that represents the latent variables in a more appropriate pattern. However, we discovered that there is an integral operation in the logarithm, which makes it impossible to expand the logarithm. This *Maximal Likelihood Estimation* (MLE) can be impossible to compute at this time. To solve this issue, a mathematical solution based on the *Evidence Lower Bound* (ELBO) is defined to provide a sound close approximation, such as

$$\begin{aligned} \mathcal{L}(q, \theta) &= \int_z q_\phi(z) \ln p_\theta(x, z) - \int_z q_\phi(z) \ln q_\phi(z) \\ &= \mathbb{E}_{z \sim q_\phi}[\ln p_\theta(x, z)] - \mathbb{E}_{z \sim q_\phi}[\ln q_\phi(z)] \end{aligned} \tag{3}$$

By taking the chain rule of

$$p(x, z) = p(z)p(x|z) = p(x)p(z|x) \tag{4}$$

we can formalize the equation by removing the joint probability $p(x, z)$ and formalize the ELBO function $\mathcal{L}(q, \theta)$ as

$$\mathcal{L}(q, \theta) = \mathbb{E}_{z \sim q_\phi}[\ln p_\theta(x, z)] - \mathbb{E}_{z \sim q_\phi}[\ln q_\phi(z)] \tag{5}$$

where the probability of $q_\phi(z)$ indicates the prior probability of any arbitrary latent variable z and finally get

$$\ell(\theta; x) = \mathcal{L}(q, \theta) + KL(q_\phi(z)||p_\theta(z|x)) \tag{6}$$

where the log likelihood of $\ell(\theta; x)$ we expected can be computed by adding the ELBO function $\mathcal{L}(q, \theta)$ to the *Kullback-Leibler Divergence* (KL Divergence) between latent prior $q_\phi(z)$ and posterior $p_\theta(z|x)$.

KL Divergence

KL Divergence is a measure that computes the amount of information lost when one uses a proposed distribution P to approximate the true distribution Q, shown as

$$D_{\mathrm{KL}}(P \parallel Q) = \sum_i P(i) \log \left(\frac{P(i)}{Q(i)} \right) \tag{7}$$

In simple words, it computes the gap between two distributions. If the prior probability is as same as the posterior probability in an ideal condition, then the second part of the equation is going to be 0, like

$$\ell(\theta; x) = \mathcal{L}(q, \theta) \tag{8}$$

Therefore, we can conclude that the goal of VAE training is to minimize the sum of reconstruction loss and the KL divergence.

Generation Validation Standards Two validation degrees from human perception and numerical metrics will be applied in this research. Computer vision processes such as image generation are used to serve human works, so the reviews from human judgements are essential and considerable. The first evaluation phase will be expanded depending on empiricism. However, the judgments made by humans are always subject and cannot be easily quantified, some more reliable standards are introduced for image generation quality examinations: the *Fréchet Inception Distance* (FID), the *Inception Score* (IS) and the *Data Composition Analysis* (DCA).

FID FID[6], the *Fréchet Inception Distance*, is a measure used to calculate the distance between the feature vectors of the real image and the generated image. We use this distance to measure the similarity between the real image and the generated image. If the FID value is smaller, the similarity is higher. The best case is FID=0 and the two images are the same.

$$\text{FID} = ||\mu_r - \mu_g||_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}) \tag{9}$$

where μ_r and μ_g are the mean vectors of the real data distribution and the generated model. Σ_r and Σ_g represent the covariance matrices of the real data distribution and the generated model. Tr denotes the trace of a matrix and $||\cdot||_2^2$ denotes the two-paradigm number of a vector.

IS IS[14], the *Inception Score*, uses an image category classifier to evaluate the quality of generated images. The image category classifier used is Inception Net-V3. This is also the origin of the name *Inception Score*. The larger the IS value, the better the model is.

$$\text{IS} = \exp\left(\mathbb{E}_{x \in G}\left[KL(p(y|x)\,||\,p(y))\right]\right) \tag{10}$$

where $x \in G$ indicates that x is an image sampled from G. $KL(p\,||\,q)$ denotes the KL-divergence between the distributions p and q. $p(y|x)$ is the conditional class distribution and $p(y)$ is the marginal class distribution. exp is there to make the values easier to compare.

DCA DCA is the third metric applied in this project, which is a straightforward strategy to compare the percentage of data categories that exist in data sets of training and generation. DCA provides insights into the data balance. It potentially indicates the image generation quality and the model learning performance.

Latent Space Modeling To model the latent space and facilitate further research processes. We first extract the latent space created by the encoders from pre-trained *VAEs* models in both $2D$ and $3D$. Then we apply the clustering approach to formalize the in-space data distribution. Some specific nodes will be sampled from the margins that exist in between data clusters to investigate what happened in the out-of-scope cases (i.e., scopes that do not belong to any single data cluster in the GMM or Gaussian distribution in the latent space). The in-cluster cases also attracted our interest, especially cases close to the center of the latent space. The central area of GMM sees a staggered part with many

irregular clusters interacting with each other, which can lead to some potential dysfunctional issues. By decoding the in-space data points and identifying their relationship with the image expression, some insights will be gained about the dysfunctional conditions.

GMM To research the capability of the deep generative models, we must have an all-around recognition and understanding of the generation procedures. To facilitate this process, the clustering algorithm is established to model the "learned information" provided by the neural network encoders. Concerning the data assumption of the Gaussian distribution, the method of *Gaussian Mixture Model* (GMM) is selected to reorganize nodes and model the latent variable space formulated. GMM is a probabilistic clustering model that has all data points generated from a mixture of a finite number of Gaussian distributions. Each composition Gaussian distribution owns its corresponding mean and covariance values, which can be present as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{11}$$

EM where $p(\mathbf{x})$ indicates the probability density function of a given data point x, K presents the total number of the Gaussian distributions, π_k provides the mixture weights for each Gaussian distribution, and $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ provides insights about each Gaussian distribution k with the relevant vector parameters of mean $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. GMM parameters of means, covariances, and mixture weights are commonly assessed by the *Expectation-Maximization* (EM) method to tune. The EM algorithm is an iterative method to find parameters maximum likelihood estimates (MLE) in learning models, relying on unobserved latent variables. It switches between two phases: the *expectation step* (E-step)

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}}[\log L(\theta; X, Z)] \tag{12}$$

which creates an expectation function $\mathbb{E}$ for the likelihood evaluation based on the existing parameter estimations $L(\theta; X, Z)$; and the *maximization step* (M-step)

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}) \tag{13}$$

which computes the latest parameters of $\theta^{(t+1)}$ that maximizes the expected log-likelihood $Q(\theta|\theta^{(t)})$ acquired from the previous step. The EM algorithm contributes to maximizing the data generation likelihood based on an arbitrary model. Distinct from some "hard clustering" methods such as *k-means*, the clustering approach from GMM is known as a kind of "soft clustering" approach which delivers a probability for each data point with respect to each possible category. By contrast, the flexible GMM allows for a more nuanced cluster assignment.

The illustration of the GMM distribution over the latent space of VAE (2D) can be found in the Figure 2. The "blue points" contained in the background indicate each feature point marked in the latent space and the black contour

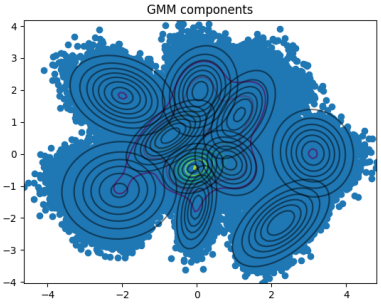

Fig. 2: The diagram provides the visualization of the GMM space, the space is mapped by the feature points present in the latent space of VAE (2D). “Contour lines" of different colours represent different probability density levels.

shows the distribution of each composition Gaussian distribution contained in the GMM space. “Contour lines" of different colours represent different probability density levels. The closer the contour line is to the center, the higher the probability density; the farther the contour line is from the center, the lower the probability density. The “black lines" are specially used to mark each Gaussian component’s contour line, indicating each component’s probability density distribution. These lines are plotted on top of coloured probability contours to show the position and shape of the individual components clearly.

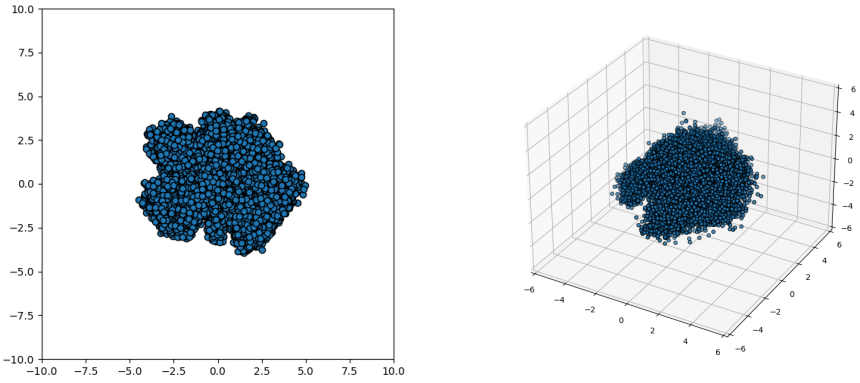

Fig. 3: The left-hand-side diagram shows a latent space created by a 2D VAE model after 100 epochs of learning, and the right-hand-side diagram illustrates the latent space under the same learning setting but in 3D.

In addition to the established GMM, the visualization of latent space itself can also be processed as illustrated in Figure 3 for both 2D and 3D cases. Some clusters can be found in the diagrams which indicate different data features learned by the encoding phase of the deep generative model of VAE. Furthermore, the training data set can also be mapped into the latent space as shown in Figure 8. The difference in node colours indicates distinct category labels they own. Ideally, the model should be able to learn an individual cluster for every single label category (i.e., each cluster should be recognized and shown in a single colour).

3.2 Marginal Dysfunction

In this part, we will explore some in-depth reasons for the “dysfunctional cases" related to the marginal area of the latent space. We will first locate some potential marginal areas manually and then sample data points from these scopes obeying the Gaussian distributions. Some batches of sampling will be processed and the results will be evaluated based on two main metrics introduced to confirm the average performance.

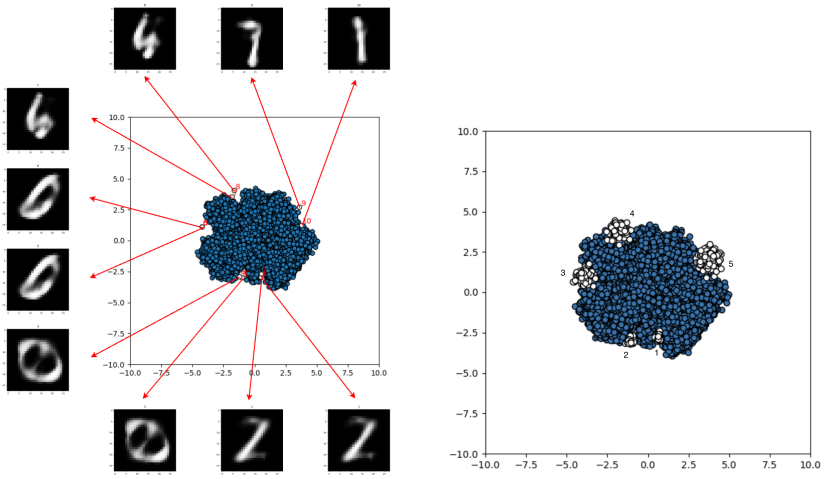

Fig. 4: **Left Subplot**: This diagram indicates five marginal “sampling areas" selected around the latent space of VAE (2D). There are two data generation points picked in each margin for research and reexamination; **Right Subplot**: This diagram illustrates the target samples selected from each marginal sampling area from 1 to 5 in the latent space of VAE (2D). Each “white node" indicates one generated data point acquired. These data points will be collected and decoded into images for further analysis.

Marginal Batch Sampling Based on empiricism and manual reexaminations, we labelled five "sampling areas" (see Figure 4 left subplot) around the expression of the latent space. Because each image is assumed to obey a *Gaussian distribution*, the sampling process will also follow the random conditions that came with the *Gaussian distribution*. This approach also guarantees the randomness of the sampling process and the diversity of the data generated. Each time we generate the sampling data points based on the distribution

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{14}$$

concerning the assumption of *Gaussian distributions*. The mean value μ is located manually to the sampling area and the variance value σ^2 (defined by the standard deviation σ) is carefully chosen concerning the size of the gap between the cluster margins.

"Batch sampling" is taken from each target area (see Figure 4 right subplot). For each round of sampling, there are 500 marginal data generation points selected and decoded with 3 rounds in total, then we acquire the average outcomes for estimation.

Decoding Evaluation After taking the "batch sampling", we will process all these images through the decoding phase to request the generated images. The image estimation follows the strategy explained in section 3.1. We have a human perception-based check about their quality and the numerical benchmarks FID and IS scores to be computed between the "Ground Truth Set" and the "Generated Data Set" for more object indications.

3.3 In-cluster Dysfunction

Some dysfunction cases found in the central area also attracted our attention. For instance, we would like to know why the image generated from the data point 5 (i.e., the point located at the center of the GMM distribution) in Figure 5 is even more vague and poor compared to some out-of-scope nodes such as data point 4 or 2. This study may help us to identify how the "dysfunctional" conditions occurred depending on the variational features of the model VAEs.

In-cluster Batch Sampling Similar to what we did in the last phase, more nodes will be sampled in this section but from more central areas of the VAE latent space. The task of this section is to sample some nodes from each identified cluster, so we can attempt to explain or recognize what detail features each specific composition distribution of the GMM recognized. The sampling will be processed by steps of "coordination selection" and "data decoding".

To facilitate the procedure, data points will be selected and taken from the latent spaces created by VAE (2D) and VAE (3D) models after the 100-epochs training process. For the step of "coordination selection", we will manually locate the position and scope of each cluster and test some representative nodes from

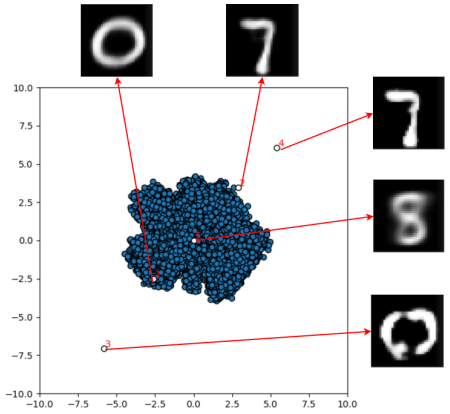

Fig. 5: The diagram illustrates how the sampling works on the latent space of VAE (2D). There are five sample nodes picked from the space with different outcomes after decoding.

each. The test outcomes can be found in Figure 6. The insignificant differences between these in-cluster nodes will be ignored and the main features learned will be concentrated.

This information will be useful as we can have a better understanding of the function of each data distribution contained in the GMM. Although not all features are recognized clearly, We roughly identify the key features learned in most of the clusters that exist in the latent spaces from both the 2D and 3D VAE models. The label indications can be found in the diagram Figure 8

For each data distribution of the GMM (i.e., each data cluster in the latent space), a feature label is assigned based on the indications from training data mapping and some practical examinations. The visualization results see clear distribution differences via the latent space of VAE (2D) and VAE (3D). However, due to the similarity in the model approach and the identification in data training, both latent space in 2D and 3D present similar cluster positions. For example, we can see that the cluster with label 4 is always located in-between the cluster with label 6 and the cluster with label 9 in both dimensions. The learning performance is also distinguishable like the cluster with the label 8 can be identified clearly in the 3-dimensional case but vague and chaotic in the 2D space.

Besides, more consequential in-cluster point generations are decoded and compared to some similar instances we found in the test data set, these provide us more evidence to prove the constant relationship between the latent space assignments and the outcome generations. More relevant issues and some potential explanations will be delivered in the later section of **??**.

Random Batch Decoding The connection between the in-space cluster recognition and the performance of the image generated is a considerable study topic

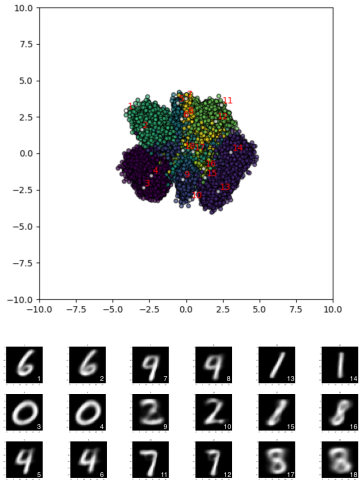

Fig. 6: The diagram illustrates how the in-cluster sampling works on the latent space of VAE (2D). Two sample nodes are picked from each cluster of the GMM with distinct decoded image outcomes listed below.

we concentrated on. Some “inspection focuses" for the following examinations are pre-defined as follows.

Inspection Focuses

For instance, there are two degrees highlighted in our comparisons from the “latent side", such as

- **L1**. the cluster **recognition clarity** in **dimensions**; and
- **L2**: the “cluster - feature" **suitability**;

and two norms for the data “generation side" like

- **G1**: the **category balancing** in image generation; and
- **G2**: the image generation **quality**;

To access more insights about the relationship, a new examination is designed to test the category-based image generation balancing and quality. By denoising a considerable amount of random noisy images, the generation preferences will be recognized and connected to the latent expression already gained previously.

For the “generation balancing test", we first examine the category distribution in the training dataset. The training data set of *MNIST* is treated as a “balanced dataset" consisting of 60000 hand-written digits, and these digits can be recognized or collected in categories from ‘0" to ‘9". The condition of the training dataset will be used as a baseline benchmark to compare with the statistical results of the generated images.

To gain a sufficient number of images for more general results, there are 30000 random image-size matrices created with random noise. These random

noise matrices will be processed by the "decoder phase" of both the models of VAE (2D) and VAE (3D) to see how these generation models learn to deal with different feature generations. We collect all 30000 images decoded from each model and assign each image to its corresponding category in all 10 categories from '0" to '9".

kNN To increase the efficiency and categorization precision, a model based on *k-Nearest Neighbors* (kNN) is prepared to handle the task. Machine learning-based classifiers perform more stably and accurately than human judgement in dealing with the high-order-of-magnitude data set. kNN algorithm is popular It operates by comparing the distances between an unknown sample $\bar{x}$ and all other learned samples $x_i \in X$ in the training dataset to determine the class of an unidentified instance by uncovering the types of the k closest data points. Based on the previous validation outcomes, the kNN model sees the most robust performance in hyperparameter case k=*3*, with the optimal performance in classifying *MNIST* images indicated by "F1-micro Score" of 97.302 and "Accuracy" of 97.892. Based on the feedback from the classifier, we made statistical comparisons on those data and found some connection between the "latent side" and the "generation side" (see section 4.2 for more details).

4 Outcome Interpretation

Depending on the research methods discussed in the previous section, the connected outcomes bring notable insights to our interpretation, recognition and judgment of the dysfunctional conditions that occurred in the image generation process. We will first present the examination products by comparing the judgement from both the *Cognitive Standard Judgment* and *Quantitative Standard Judgment*. Furthermore, the patterns recognized in the generation dysfunctions will be declared, with some theoretical explanations furnished in the last subsection.

4.1 Marginal Dysfunctional Analysis

Cognitive Standard Judgment For the analysis made by human perception and empiricism, we first compared some notable samples with original images that exist in the test cases from similar categories (see Figure 9).

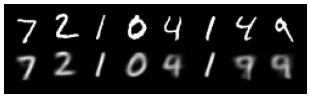

Fig. 7: The diagram compares the generated images from the marginal data points (the second row) with some corresponding original samples (the first row) found in the test set under the same categories. Both VAE (2D) and VAE (3D) present the similar dysfunctional issues.

We can see that the generated images from the marginal data points present in the second row came with much worse clarity. Both VAE (2D) and VAE (3D) present the similar dysfunctional issues. Some digital categories are even difficult to distinguish. For instance, perception judgment argues that the generation process cannot differentiate the digits “4" and “9" to a large extent.

Quantitative Standard Judgment Based on the batches of marginal instants sampled in subsection 3.2, we decode these data points generated and process the estimation of FID and IS scores between the generated image data and the original image set. The computation results for both VAE (2D) and VAE (3D) can be found in the table of Table 1.

Table 1: IS and FID of the generated sets from VAE(2D) and VAE(3D)

Generated Image Sets	IS	FID
VAE(2D) - Marginal Generations	11.284	220.916
VAE(3D) - Marginal Generations	20.884	188.870
VAE(2D) - General Generations	5.646	112.903
VAE(3D) - General Generations	6.321	107.746

The table outcomes denote a clear pattern where the VAE (3D) outperforms the VAE (2D) in generating quality and the marginal cases bring significant dysfunctional outcomes. The higher IS score for the VAE (3D) suggests that it brings better diversity in image generation. Furthermore, the lower FID score shows that these images have greater similarity to the training images, which implies better generation quality. The numerical trend suggests that the additional dimensionality in VAE(3D) offers an advantage in generating more complex and accurate representations over VAE(2D). Although the 3D-VAE shows slightly more acceptable resistance to the dysfunctional cases, the dysfunctions still exist around the margins in dimensions.

A stage conclusion can be made that both the human perception judgement and the numerical analysis support the opinion that dysfunctional conditions can appear in the marginal areas around the central part of the VAE latent space learned in dimensions.

4.2 Central Dysfunctional Analysis

Cognitive Standard Judgment The human perception analysis combined with manual spec-point sampling contributes a lot to identifying the dysfunctional cases that occurred in the central scope of the latent space. Refer to the samples we acquire in Figure 6, although most images generated indicate high quality and clear “boundary" in-between digit categories, some specific cases located in the middle of the latent space came with some confusing and invalid outcomes. For instance, the samples with IDs of “15" and “16" at the most central

area and the points of “17" and “18" nearby. These points present similar dysfunctional issues in that the model did not recognize the obvious gap between two digital sorts constructively, and the space center sees too many potential category distributions overlap with each other.

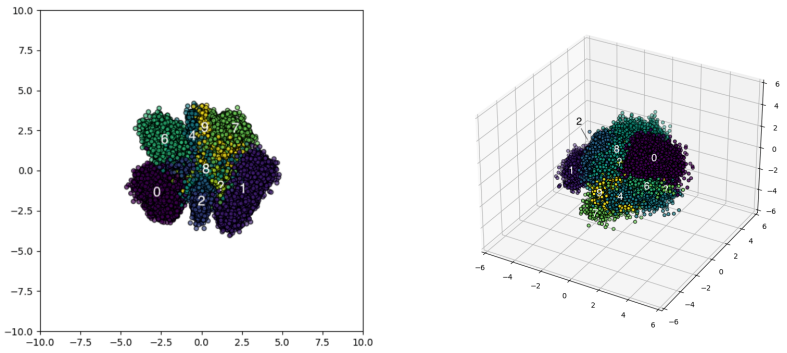

Fig. 8: These diagrams provide label indications about the features learned by each cluster in the latent space of VAE (2D, **Left Subplot**) and VAE (3D, **Right Subplot**). A few “question marks" indicate the area with unclear or hard-recognized knowledge.

Gray Area Based on the other diagram Figure 8, the most significant features for each category are labelled, except the digital categories of “3" and “5" that should appear in the center. Some samples picked from the generation set also failed the cognitive judgement: the difference between the three digits of “3", “5" and “8" cannot be identified clearly (see Figure 9). The “question marks" assigned in both subplots raise a similar concern that the central area of the latent space (p.s., in both 2D and 3D cases) presents a “gray area". The visualization of GMM we have shown previously (refer to Figure 2) also supports this opinion that the “too-overlap" state of data clusters may lead to a “feature loss" in pattern recognition.

Quantitative Standard Judgment Followed by the method described in section 3.3, the result corresponding to the generation sample sets from both the VAE (2D) and VAE (3D) are collected and analyzed. Refer to outcome data shown in Table 2, which describes the distribution of digit types across an original training set and those generated by two models, VAE(2D) and VAE(3D). The training set shows a uniform distribution with each digit type roughly representing 10% of the whole training dataset. However, the generative models deviate from this uniformity: VAE(2D) overrepresents digits “0", “3", and “4", while severely underrepresenting ‘5", ‘8", and ‘9". The VAE(3D) generated data

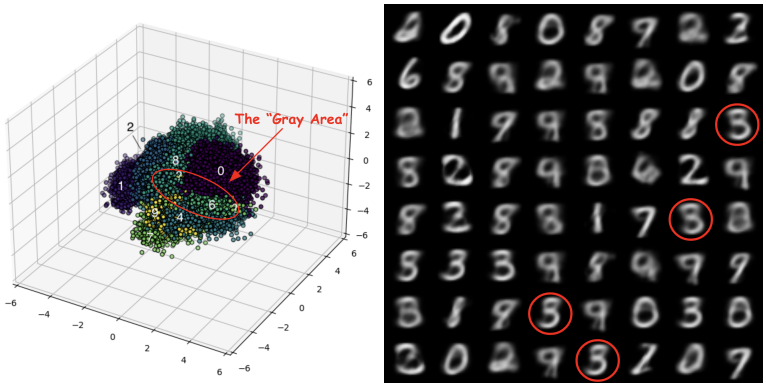

Fig. 9: **Left Subplot**: a possible illustration of the "Gray Area" in VAE (3D); **Right Subplot**: Some unclear and hard-identified cases about digits "3", "5" and "8" that found in the generated image set.

skews heavily towards "1" and '4", with these types constituting well over the expected average, but it underrepresents '2", '9", and especially "8".

Table 2: The distribution of digit types in VAE(2D) and VAE(3D)

Types of Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Sum**	**Avg**
Training Set(#)	5923	6742	5958	6131	5842	5421	5918	6265	5851	5949	60000	6000
VAE(2D) - Generated Set(#)	4428	3786	2576	6652	4175	354	3591	2008	566	1864	30000	3000
VAE(3D) - Generated Set(#)	2948	6823	1476	4253	4751	2425	2089	2681	1750	814	30000	3000
Training Set(%)	9.87%	11.24%	9.93%	10.22%	9.74%	9.03%	9.86%	10.44%	9.75%	9.92%	100%	10%
VAE(2D) - Generated Set(%)	14.76%	12.62%	8.59%	22.17%	13.92%	1.19%	11.97%	6.69%	1.89%	6.20%	100%	10%
VAE(3D) - Generated Set(%)	9.79%	22.74%	4.92%	14.18%	15.84%	8.08%	6.96%	8.94%	5.84%	2.71%	100%	10%

This table shows the results of the DCA. We detected the number and percentage of different types of digits in different sets, including the original training set and the generated set of 2D VAE and 3D VAE.

Both models exhibit challenges in capturing the balanced digit distribution of the training set, but the VAE(3D) demonstrates a slightly more balanced distribution across certain digits when compared to the VAE(2D), which has more extreme variances. The found in this part argue that the generation outcomes are "extremely unbalanced" (see Figure 10) even when perfectly balanced training sets are used. Fortunately, the issue of "unbalancing generation" diminishes with the model dimension increasing.

Additionally, we also care about the image generation quality from the in-cluster cases depending on each data category. The Table 3 compares the FID scores for individual digit classes generated by VAE (2D) and VAE (3D). A general pattern emerges where VAE (3D) outperforms VAE (2D) for most digit types, indicated by lower FID scores, which implies closer similarity to the actual digit distributions. Notably, VAE (3D) shows significantly better performance for digits "1", "4", "7", and "9", suggesting it can generate these digits with higher

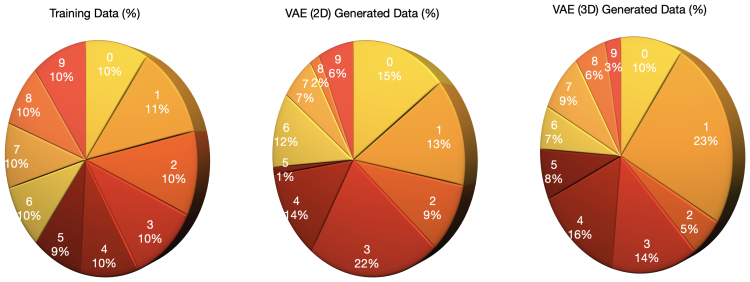

Fig. 10: The diagram visualizes the outcomes od the DCA, which compares the "balancing levels" from each dataset of the training set, generation set from VAE (2D), and the generation set from VAE (3D).

fidelity than VAE (2D). The VAE (2D) model, however, has its lowest FID score with digit class "1" and highest with digit class "7". Across both models, the digit "5" appears challenging, with higher FID scores compared to other cases. This table suggests that the 3D model generally captures the data distribution of individual digits more effectively than its 2D counterpart with more appropriate images created, but both models exhibit varying degrees of difficulty across different digit types. The analysis of generation quality and balancing for 2D and 3D VAE models reveals that while both struggle to maintain the uniform distribution of the training set, the VAE (3D) is more adept at learning and fitting the underlying data distribution.

Table 3: FID of different digits in the generated set of VAE(2D) and VAE(3D)

Type of Digits	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
VAE(2D) - Generated Set	207.4569	171.6291	210.5091	197.6481	191.3871	228.2051	251.1663	258.4876	218.7353	231.7673
VAE(3D) - Generated Set	239.6646	127.0432	175.0209	171.9411	134.1443	189.7628	201.1065	142.4139	165.2396	141.6462

We calculated the FID of each type of digit in the generated 2D VAE and 3D VAE respectively.

4.3 Pattern Recognition

Concluding our study, we've ascertained that generative models can have dysfunctional situations appear in both the marginal area and the in-cluster area. From the latent perspective, our comparison highlighted two key aspects: the recognition clarity of clusters within the latent dimensions (L1) and the suitability of clusters to their corresponding features (L2). These focal points are critical in determining the model's ability to discern and accurately map the data's inherent structure. Our visualization results and perception judgements support that both the marginal and central areas may lead to dysfunctional issues and have the central "gray area" with poor feature-fittings. By contrast, the issues of

“feature loss" are more significant in the central “gray area” due to under-fitting in a limited space dimension. The model with a higher latent space dimension may have more appropriate outcomes and higher resistance to “feature loss".

On the generation side, we investigated the models based on category balancing in image generation (G1) and the overall quality of the generated images (G2). Even though we found that using a balanced training data set may unavoidably lead to unbalanced image generations, our findings reveal that VAE (3D) outperforms its 2D counterpart in both considerations of generation quality and balancing. The 3D model demonstrates superior cluster recognition clarity and feature suitability in both the marginal and central samplings, leading to more distinct and representative latent representations. Concurrently, it excels in producing a more balanced assortment of categories and higher-quality images, as evidenced by lower FID scores and a closer resemblance to the training data distribution. These studies deliver notable insights for future improvement and fixing over dysfunctional image generations.

4.4 Explanation Discussion

a. Fitting and Recognizing Issues with generative models, such as bland outputs or unclear category representation, often boil down to the model not being complex enough to handle the data intricacies, which is a problem known as “under-fitting". This can happen when the data at hand doesn't present enough variety, or when the model training is too basic.

To fix this, strengthening up the dataset can be a good choice: getting more data with more variety. Allowing the generation model to have longer training may also help to learn the ropes. In this way, much clearer “boundaries" between the data distributions can be identified, so the dysfunctional issues due to “too-overlap segments" can be reduced. Hyperparameter fine-tuning may also be a considerable measure in enhancing information fitting performance and bringing better pattern recognition ability against dysfunction conditions.

b. Information and Dimension For the consideration of information and dimension, our research suggests a compelling relationship between the dimensionality of the space in which the model operates and the clarity of feature clusters it produces. Higher-dimensional spaces can preserve more features from the input information to present a robust resistance to feature loss with more balanced outputs. The reason can be that a higher-dimensional space comes with a sparser space where each information cluster can spread out. This setting lowers the overlapping chance thus bringing clearer boundaries between the feature category.

Besides, the "roomier" nature of such spaces enables more organized image generations and reduces the between-clusters dysfunctional conditions. Furthermore, the model with expansive spaces can create a multi-faceted understanding of each feature cluster, which helps capture some subtle nuances that may be lost in low-dimensional expressions. The higher dimensional operation also serves

more precise and nuanced mappings between features and respective clusters, potentially resulting in a generative process that better reveals the training data diversity. Ideally, the latent space of VAE should be continuous and without gaps, but in practice, VAE may not fully cover the latent space, resulting in uneven quality of the generated information.

c. Divergence and Assumption In addition to the explanations from learning and expression, we can also dig into some in-depth design reasons that lead to the dysfunction conditions. Referring to the inference processed before and the final learning equation of , the goal of VAE training is to minimize the sum of the "reconstruction loss" and the "KL divergence". The smaller the "reconstruction error", the closer the data generated by the model is to the real data. By contrast, "KL divergence" measures how regular the latent space is and how it follows the prior distribution.

Vanishing KL Div. The first notable issue exists in the control of "reconstruction Loss" and "KL divergence". As Shao et al. argued, "if the KL-divergence is too high, it would affect the accuracy of generated samples. If it is too low, output diversity is reduced." [15] In the VAE training and optimization processes, the reconstruction loss may dominate the overall error calculated. If the model weights too heavily on minimizing the reconstruction error, this may cause the KL divergence term to approach zero, known as the problem of "vanishing KL divergence".

In this case, VAE may no longer wish the latent space to approach the prior distribution, causing the model to lose the "generation ability" and "learn too well" about the noisy information that came from the training set. Besides, due to the same reason, "vanishing KL divergence" may also cause the degradation of the latent space to lose its desired smoothness and continuous properties. This regularization term of "KL divergence" may inhibit some feature expressions and constructions in the latent space significantly and finally lead to dysfunctions in image generation.

MSE Some other design issues may come from the selection of error functions and the assumption of images. "Reconstruction losses" typically use pixel-level *mean squared error* (MSE), but this error metric does not always fit the human judgement of graphical quality. Additionally, the naive assumption for images supported by the discrete Gaussian distribution may also lead to some problems. Although this convenient setting helps to formalize the process of image processing it brings unavoidable feature loss which may also lead to dysfunctional conditions happening in image generation.

5 Conclusion

This paper demonstrates the studies for recognizing and interpreting dysfunctional cases in image generation. By exploring the marginal and central area of the latent space in VAE, some dysfunctions like "latent feature loss", "unbalancing generation" and "vanishing divergence" are appropriately recognized

and specified during the image generation process. Our experiment provides sufficient and reasonable interpretations for dysfunctions from both degrees of the "implicit latent space" and "explicit data generation."

Some possible interpretations and explanations about the dysfunctional cases are provided in three dimensions data fitting, space expression, and mathematical reasoning. Under-fitting issues may be a considered problem which restricts the model from learning clear image generation; The “gray area" in the center of latent space may be caused by “low dimensional feature loss" in information and “too-overlap" in space distribution; For the "unbalancing generation" problem, although a balanced data set may lead to unbalanced generation, increasing dimensions help diminish this problem; Also, the mathematical inference argue that the "vanishing KL divergence" may lead to latent space degradation with generation properties loss, thus causing more dysfunctions. All these outcomes provide notable insights for future enhancement in image generation and raise possible solutions to handle the existing problems.

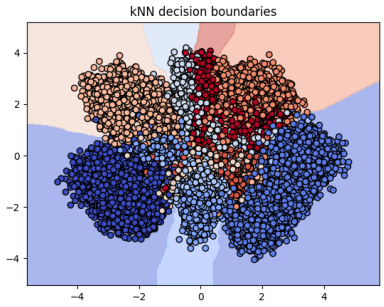

Fig. 11: The diagram visualizes the latent space of VAE (2D) covered by the classification of kNN.

We believe the issue of dysfunctional image generation is an intriguing investigation which could be usefully explored in further research about deep generative models. Some interesting innovation points found during the dysfunctional research may be worth mentioning. For instance, we imagine a possible VAE variant that can apply the kNN model to regularize its latent space in training (see Figure 11). As we discussed before, the latent space with clear “decision boundaries" will contribute a lot in controlling the generation loss with more robust resistance to dysfunctions. The kNN is also known as a spatial classification method, which can be easily applied to the latent space inside generative models. It typically creates semi-linear boundaries to split data collections with a reasonable distance between each other, which perfectly fits the expectation for a satisfactory latent space distribution. More connected interpretational research and recognizant applications are desired in response to the issues of dysfunctional image generation.

References

1. Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models, 2021.
2. Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks, 2019.
3. Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S. Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *ArXiv*, abs/2303.04226, 2023.
4. Tong Che, Yanran Li, Athul Paul Jacob, Yoshua Bengio, and Wenjie Li. Mode regularized generative adversarial networks, 2017.
5. Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Scholkopf, and Alex Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, 2012.
6. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.
7. G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.
8. Tianyang Hu, Fei Chen, Haonan Wang, Jiawei Li, Wenjia Wang, Jiacheng Sun, and Zhenguo Li. Complexity matters: Rethinking the latent space for generative modeling, 2023.
9. Hui Jiang. A latent space theory for emergent abilities in large language models, 2023.
10. Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.
11. David Lopez-Paz and Maxime Oquab. Revisiting classifier two-sample tests, 2018.
12. Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Refusion: Enabling large-size realistic image restoration with latent-space diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 1680–1691, June 2023.
13. Sudipto Mukherjee, Himanshu Asnani, Eugene Lin, and Sreeram Kannan. Clustergan: Latent space clustering in generative adversarial networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):4610–4617, Jul. 2019.
14. Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.
15. Huajie Shao, Shuochao Yao, Dachun Sun, Aston Zhang, Shengzhong Liu, Dongxin Liu, Jun Wang, and Tarek Abdelzaher. Controlvae: Controllable variational autoencoder, 2020.