EVALUATING GENERATIVE MODELS FOR MEDICAL IMAGES FROM MEDMNIST

Xingyu Zhang, Zeying Huang, Scott Sun

ABSTRACT

Conditional Generative Adversarial Network (cGAN), affording the control of synthesized sample's classes, is an improved version of GAN. In this study, we employ Fréchet Inception Distance (FID) to assess the quality of synthetic medical images generated by cGANs. We contrast cGANs implemented in the Wasserstein architecture with their vanilla counterparts. Furthermore, we investigate whether encoding class labels with embeddings of various sizes can outperform using one-coding representations.

1 INTRODUCTION

Generative Adversarial Network (GAN) (Goodfellow et al., 2014) is a powerful class of generative neural network model. The generator module learns to synthesize fake data based on random noise through a "fight" against the discriminator module. Conditional GAN (cGAN) (Mirza & Osindero, 2014) allows for controllable generation, where we can provide the model additional information to synthesize data of our interest. Nevertheless, Nevertheless, the inherent minimax optimization nature complicates the training process. Therefore, the synthetic data may be in poor quality. Wasserstein GAN (WGAN) (Arjovsky et al., 2017) is well-known for effectively address common problems in the vanilla form, so we extend cGAN to the Wasserstein framework and construct conditional WGAN (cWGAN).

Our primary aim in this work is to generate MedMNIST (Yang et al., 2023) images using cGANs under both vanilla and Wasserstein frameworks. Subsequently, we will evaluate and compare various models by visually assessing their generated images and computing the unconditional/conditional FID scores (Heusel et al., 2017).

2 RELATED WORKS

2.1 MEDMNIST DATASETS

MedMNIST is a large-scale MNIST-like collection of biomedical images, where all 2D images are in the shape of 28×28 (Yang et al., 2023). In our work, we study on three datasets: BloodMNIST, PneumoniaMNIST, OrganAMNIST. Table 1 provides basic information about the datasets.

Table 1: MedMNIST Datasets

Dataset	Modality Description	#channels	#classes	#Train/Valid/Test
BloodMNIST	Blood Cell Microscope	3	8	11,959 / 1,712 / 3,421
PneumoniaMNIST	Chest X-Ray	1	2	4,708 / 524 / 624
OrganAMNIS	Abdominal CT	1	11	34,581 / 6,491 / 17,778

2.2 CGAN

Conditional generation behaves like picking a item in a vending machine. In cGAN, both the generator and discriminator are conditioned on extra auxiliary information (Mirza & Osindero, 2014). In our case, the extra information is class label, denoted by y, so the value function is Eq 1.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x|y)] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z|y)))] \tag{1}$$

Using the $-\log D$ trick proposed by Goodfellow et al. (2014), we can rewrite the minimax problem into alternating optimization given by Eq 2.

$$\begin{cases} \max_D \left(\mathbb{E}_{x\sim p_{\text{data}}}[\log(D(x|y))] - \mathbb{E}_{z\sim p_z}[\log(1-D(G(z|y)))] \right) \\ \min_G \left(-\mathbb{E}_{z\sim p_z}[D(G(z|y))] \right) \end{cases} \tag{2}$$

2.3 WGAN and Gradient Penalty

In Wasserstein GAN (WGAN), the final Sigmoid activation function is removed from the discriminator. As a result, the discriminator becomes a "critic" that rate images' degree of realness with scores in $\mathbb{R}$ rather than classifying if it is real or fake.

WGAN stabilizes optimization, mitigating mode collapse and discriminator saturation, thereby enhancing diversity and fidelity in generated samples (Arjovsky et al., 2017). Arjovsky et al. (2017) enforces Lipschitz constraint by weight clipping. It is straightforward but sensitive to choice of the clipping threshold, c. A large c may lead to sluggish critic convergence, and a small c may lead to vanishing gradients (Arjovsky et al., 2017).

Gradient penalty (WGAN-GP) (Gulrajani et al., 2017) frees us from hyper-parameter tuning by merging the Lipschitz constraint into the objective function as a regularization penalty. The updated alternating optimization is given by Eq 3

$$\begin{cases} \min_D \left(\mathbb{E}_{z\sim p_z}[D(G(z))] - \mathbb{E}_{x\sim p_{\text{data}}}[D(x)] + \lambda \mathbb{E}_{\hat{x}\sim p_{\hat{x}}}\left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2 \right] \right) \\ \min_G \left(-\mathbb{E}_{z\sim p_z}[D(G(z))] \right) \end{cases} \tag{3}$$

In Eq 3, $\hat{x} = tG(z) + (1-t)x$, $t \sim \text{Uniform}(0,1)$, so that $p_{\hat{x}}$ samples uniformly along straight lines between pairs of points from p_{data} and $p_{G(z)}$ (Gulrajani et al., 2017).

2.4 FID

A well-trained CNN transforms real or synthetic samples into embedding vectors, intermediate activation scores before the final FC layer with Softmax activation. The embedding size may vary based on the CNN architecture (e.g., 2048 for a ResNet50 model). Using the CNN model, we can calculate embeddings for both real samples and generated samples, respectively. FID is a metric to compare these two sets of embeddings (Heusel et al., 2017). It is calculated by assume multivariate Gaussian distributions, which is shown in Eq 4.

$$\text{FID} = \|\mu_{\text{real}} - \mu_{\text{generated}}\| + \text{tr}\left(\Sigma_{\text{real}} + \Sigma_{\text{generated}} - 2(\Sigma_{\text{real}}\Sigma_{\text{generated}})^{1/2} \right) \tag{4}$$

3 Details of the project

3.1 cGAN & cWGAN

Our initial exploration focused on determining how best to integrate the label information with image input or latent noise. We implemented two different methods to encode class label information: embedding encoding and one-hot encoding. Subsequently, the challenge was to effectively combine these two modalities. A straightforward approach was to use FC layers and reshaping to blender different information together. Figure 1 is a sketch of this architecture [1].

[1] It took me a while to create these images in MS PowerPoint

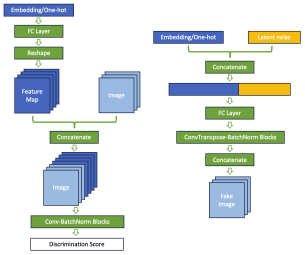

Figure 1: Naive approach.

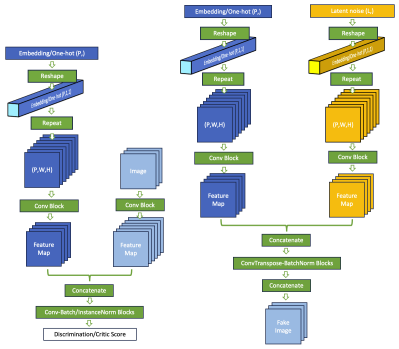

Figure 2: Improved approach.

However, the generator in this architecture was prone to have saturation problems in our empirical experiments: the discriminator quickly converged while generated samples had mode collapse and were in poor quality. Hence, we did not extend it to the Wasserstein framework.

Then, we explored another design proposed by Zhou (2022) in her lecture on Coursera. We internalized the idea of transforming encoding vectors to tensor with constant layers. Figure 2 is a sketch of the new structure, which is also provided in our finalized code. We then simply ran a few experiments on the new architecture under the vanilla setting. The results turned to be reasonable at a glance, so we extended the structure to the Wasserstein framework.

After the sanity check, we extend the revised architecture to WGAN-GP framework and compared this Wasserstein design to its vanilla counterpart. The underlying convolutional and batch-norm building blocks are almost the same for both type of structures except that in the discriminator/critic of Wassertein cGAN (cWGAN) we replaced the commonly used batch normalization with instance normalization. As Gulrajani et al. (2017) explained, the gradient penalty was applied to each sample independently not to a batch on average. Their original work utilized layer normalization, but we chose instance normalization as it was more direct. During the training of cWGAN, we set $\lambda = 10$ as the penalty coefficient and $n_{\text{critic}} = 5$ as the number of iterations in the inner training loop for the discriminator/critic. To maximize data utilization, we trained all cGAN models using both the train and test splits of the datasets.

3.2 ResNet50 and FID

We employed ResNet50 as the CNN classifier to extract image embeddings. For each dataset, we independently trained a ResNet50 and allowed it to achieve a sufficient accuracy score on the test data to ensure the image embeddings were useful. Accuracy scores were summarized in Table 2. Data in the train and test splits were used for training, and the valid split was used for testing.

Table 2: ResNet50 performance

Dataset	Testing accuracy
BloodMNIST	0.935
PneumoniaMNIST	0.965
OrganAMNIS	0.971

The calculation of FID followed Eq. (4). We fixed the number of generated samples at 10,000 and used `sqrtm` from `scipy.linalg` to compute $(\Sigma_{\text{real}}\Sigma_{\text{generated}})^{1/2}$. However, `sqrtm` was performed on the CPU and consumed 20-30 seconds on our machines. As the formula only required trace, explicitly computing the entire square-root matrices might be unnecessary and inefficient.

4 CONTRIBUTION OF EACH MEMBER OF THE TEAM

- Scott Sun: developed classes and methods for cGAN, cWGAN, CNN, and FID; tested code; trained CNNs; edited report.
- Jeff Huang: developed and trained vanilla cGAN experiment; drew the FID experiment figures; edited report.
- Xingyu Zhang: developed and trained cWGAN experiments and computed FID results for each cWGAN model; generated fake image sets using all cGAN models; edited report.

5 EXPERIMENTAL RESULTS

This section focuses on providing a comparative analysis of synthetic blood cell microscope images (Blood data set) generated by two variants of cGAN — vanilla cGAN and cWGAN — across different training epochs and encoding strategies. Results and analysis for Abdominal CT (OrganA data set) chest X-ray images (Pneumonia data set) are provided in Appendix A for saving space. Both visual assessments and FID results are analyzed to compare the performance of the employed cGAN architectures, and further explore their capabilities and limitations on biomedical images.

5.1 GENERATED SAMPLES

Overall $2 \times 2 \times 4$ experiments are conducted on three different data sets while the first 2 represents two different GAN architectures, the second 2 includes two training epoch settings (10&100), and the 4 includes one-hot, embedding size 4, embedding size 8, and embedding size 4. To facilitate comparison, the generated samples are organized as shown in the Table 3.

Table 3: Layout of the results (for each class in each data set).

Real Data	16			
Vanilla cGAN-epoch10	4	4	4	4
Vanilla cGAN-epoch100	4	4	4	4
cWGAN-epoch10	4	4	4	4
cWGAN-epoch100	4	4	4	4
	one-hot	embedding 4	embedding 8	embedding 32

As you can see in Figure 3, the generated blood cell images using vanilla cGAN and cWGAN models closely resemble real data, especially after 100 training epochs and with increased embedding sizes. The models, particularly after extended training, excel in capturing the essential details of blood cells, including distinct cell boundaries and critical internal structures vital for precise classification—such as the granularity seen in eosinophils or the characteristic shape of lymphocytes. The role of embedding size is also significant; While the one-hot encoding provides a good foundational representation, the increase in embedding size from 4, 8, to 32 allows for a richer and more nuanced portrayal of the cells. Therefore, the cWGAN images, especially with larger embeddings exhibit a high degree of clarity and detail, indicating the models' advanced learning of data complexity and variability in cell types.

The Figure 7 displays a variety of synthetic organ images generated by different models. From a non-professional standpoint, the visual assessment of these images reveals that the synthetic data—represented in all classes except the lung—is relatively challenging to differentiate and comprehend. To be more specific, images of organs such as the bladder, left and right femurs, heart, left and right kidneys, liver, pancreas, and spleen demonstrate a high degree of complexity, which may explain the significant differences between the generated data and real data. In contrast, images of the lungs (left and right) are more recognizable, possibly due to their unique texture patterns and contrasts that the generative models seem to capture and reproduce more effectively. In addition, generated lung data trained 100 epochs with one-hot encoding or 32 embeddings performs much better than other models.

The generated images in figure 8, particularly those depicting pneumonia, are notable for their clarity in structure and the prominent white areas indicative of the condition. This suggests that the

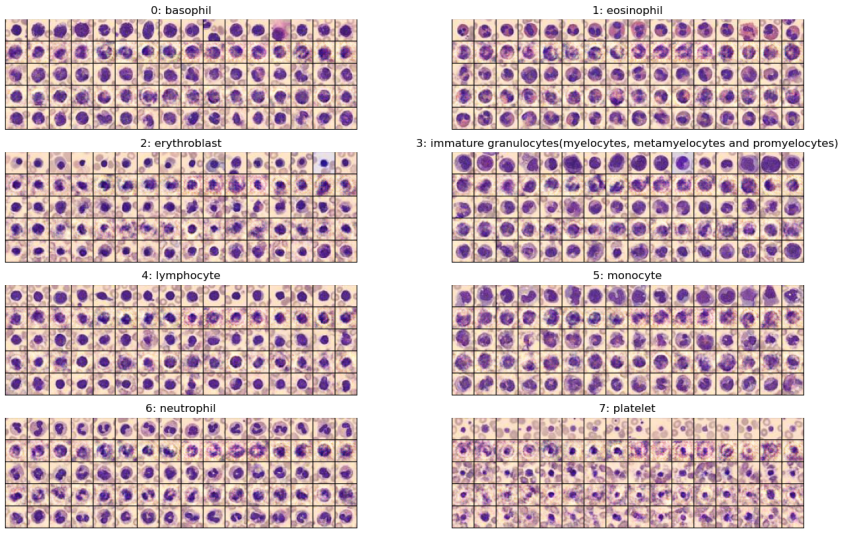

Figure 3: Generate results for Blood data set

generators are effectively learning the distinctive features of the chest, including pathological signatures specific to pneumonia. Images generated with only 10 epochs of training display noticeable 'noise', a trait more evident in those produced by Vanilla cGAN, suggesting inadequate training to fully grasp the basic structure of the chest in the actual data. Conversely, the quality of images from different cGAN architectures after 100 epochs of training is significantly enhanced, showing reduced noise and more distinct features. Furthermore, increasing the embedding size from one-hot to larger dimensions—4, 8, and 32—leads to a perceptible improvement in texture and overall image quality.

5.2 Analysis of FID score on Blood Data set

Figure 4 shows that cWGAN demonstrates a more consistent generation quality across different blood cell types, as indicated by a narrower spread of FID scores, especially with an embedding size of 32. This consistency could be better than vanilla cGAN due to its stable training and better convergence properties, leading to uniform FID scores across classes and embedding sizes.

Secondly, The impact of embedding size on FID scores varies across specific blood cell types. Eosinophils and monocytes show high sensitivity to embedding size changes, while lymphocytes and neutrophils show less variance. Notably, basophils and platelets improve at an embedding size of 8 but worsen at 32, suggesting an optimal embedding size for these classes. This indicates that each blood cell type has unique features differently captured by the cGANs based on the embedding size.

Moreover, cWGAN, using One-Hot encoding, significantly outperforms vanilla cGAN in FID scores, highlighting the advantages of discrete and unambiguous representation provided by One-Hot encoding.

Finally, the change in embedding size from 4 to 8 shows a more pronounced effect than from 4 to 32, suggesting diminishing returns with increasing embedding size.

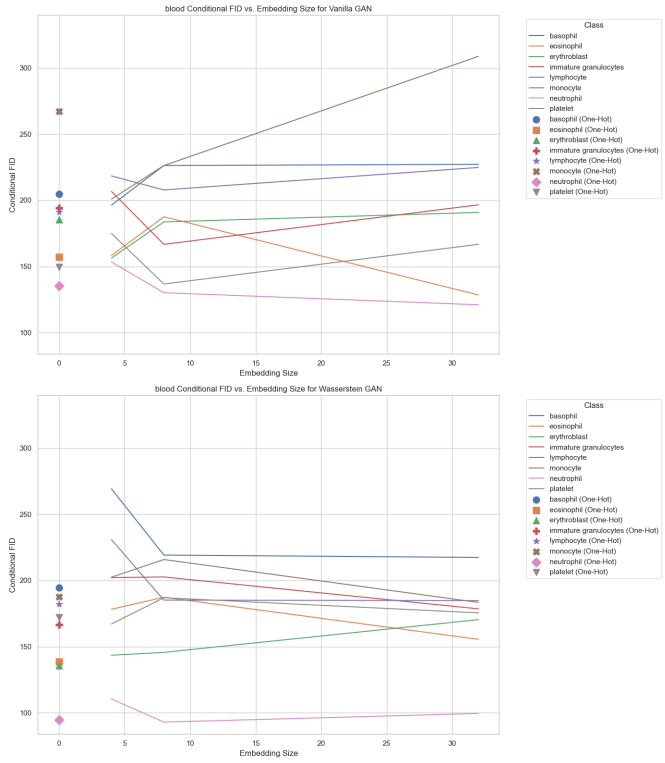

Figure 4: Conditional FID score vs. Embedding size for Blood data set

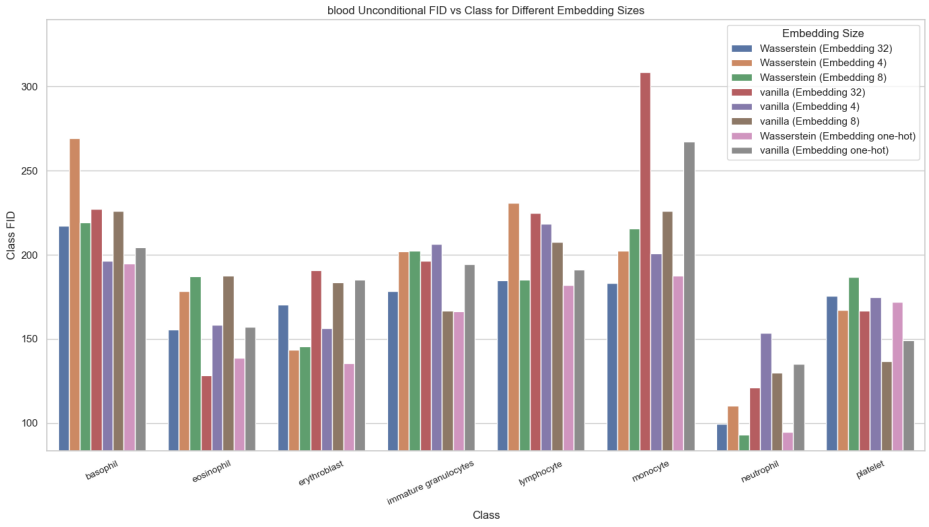

Figure 5: Conditional FID score vs. Embedding size for Blood data set

The Figure 5 shows that cWGAN generally outperforms vanilla cGAN across various classes and embedding sizes, showing more consistent and superior performance. This is evident in its handling of different blood cell types, where it maintains stable FID scores, indicating better overall model effectiveness.

What is more, sensitivity to embedding size varies among classes. Some, like eosinophils and monocytes, show significant FID score changes with different embedding sizes, indicating a strong

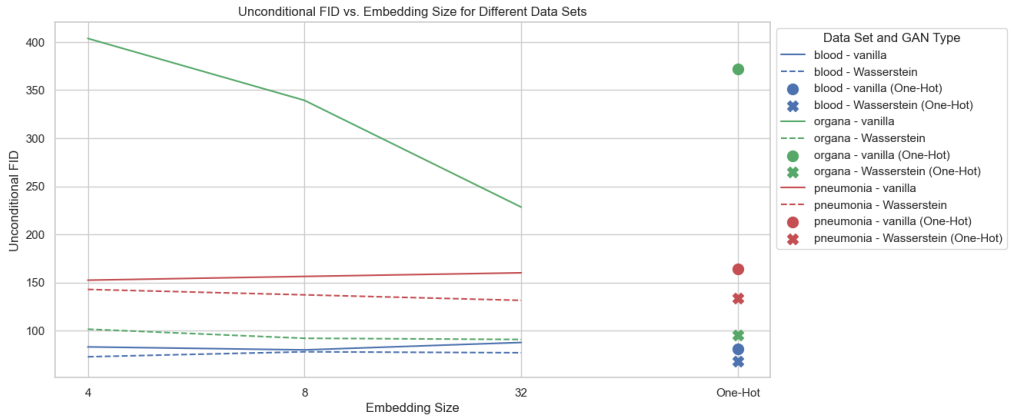

Figure 6: Conditional FID score vs. Embedding size for Blood data set

dependence on model capacity. Others, like neutrophils, demonstrate improved performance with larger embeddings across both models.

Additionally, one-hot encoding often enhances performance, particularly in vanilla cGAN. This suggests that discrete representation may be more effective for capturing complex features in certain classes than continuous embeddings.

In general, the analysis suggests that different blood cell classes require varied model capacities and representations for optimal image generation. The effectiveness of cWGAN across a range of classes and conditions highlights its robustness and adaptability in handling diverse feature sets.

5.3 Analysis of FID score on Different Data set

The Figure 6 shows that cWGANs consistently exhibit lower FID scores than vanilla cGANs across all datasets, indicating their superior capability to produce higher-quality images.

The pneumonia dataset presents higher FID scores for both cGAN types, suggesting a lower quality or less representative generated images than the blood and organ datasets—the latter two show similar FID scores, indicating comparable performance levels.

cGANs using embeddings generally have lower FID scores for the blood and organ datasets than those with one-hot encoding, implying that embeddings might capture more detailed features. Conversely, one-hot encoding yields lower FID scores for the pneumonia dataset, an exception suggesting its effectiveness for this specific data type.

The impact of changing embedding size on FID scores is less significant than expected. In Wasserstein models, an increase in embedding size tends to improve performance across most datasets. For vanilla cGANs, noticeable improvement with larger embeddings is mainly seen in the organ dataset, while smaller embeddings perform better in other cases. This may reflect vanilla cGANs' varied dependency on feature representation complexity across different datasets.

In summary, the type of cGAN model, choice of encoding method, and embedding size significantly influence image generation quality, interacting differently across various datasets. cWGANs demonstrate better stability and image quality, while the effectiveness of encoding methods and embedding sizes varies depending on the dataset.

6 Concluding Remarks

The experimental results demonstrate the varying performance of vanilla cGANs and cWGANs across different datasets and label encoding strategies in medical imaging applications. cWGANs, known for their stable training and convergence properties, consistently show narrower spreads of FID scores, indicating more stable generation quality across various blood cell types. This stability

is maintained even with larger embedding sizes, suggesting better performance consistency than vanilla cGANs.

In the context of specific blood cell classes, such as eosinophils and monocytes, the cWGAN displays less variance in FID scores with changes in embedding size, implying a higher sensitivity to embedding capacity. Conversely, lymphocytes and neutrophils exhibit lower variances, indicating a lower sensitivity to embedding size changes. The performance across classes suggests that each blood cell type has unique feature sets that are captured variably by cGANs, depending on the embedding size. Moreover, implementing One-Hot encoding in cWGANs results in significantly better FID scores, highlighting the advantage of discrete and unambiguous representation in capturing complex features.

Similarly, for the organ dataset, the cWGAN shows more consistent performance across different organ classes, benefiting from the distinct and categorical nature of the One-Hot encoding. The sensitivity of different organ classes to the model type and embedding size highlights the need for specific model capacities and representations for optimal image generation.

Both types of cGANs benefit from an increase in embedding size for the pneumonia dataset, but the effect is more pronounced in the vanilla cGAN. The initial increase from an embedding size of 4 to 8 results in a significant decrease in FID for vanilla cGAN, but further increases show diminishing returns.

Overall, the cWGANs outperform vanilla cGANs across various classes and embedding sizes, mainly using One-Hot encoding. Moreover, the failure of the naive approach indicates that we need to avoid the usage of FC layers and reshaping for input treatment in a deep convolutional structure. The analysis reveals the importance of choosing the appropriate model and embedding method for each class and the potential benefits of adaptive or class-specific cGAN architectures for medical image generation. The results also suggest that while increasing the embedding size can improve performance, there is a point of diminishing returns beyond which further increases do not translate to proportional gains in image quality.

In summary, the FID scores improve with increased embedding sizes and training epochs and performed better on cWGAN models, aligning with visual assessments that underscore the importance of extensive training and embedding complexity in generating high-quality synthetic images via cGANs. The produced images of various blood cells and pneumonia show a strong visual correlation to authentic images, demonstrating that cWGANs are capable of learning and mimicking detailed characteristics of different cell types and distinguishing features of normal and pneumonia-affected chests, suggesting their potential utility in expanding medical imaging datasets. Nevertheless, the resemblance between generated organ images and their real counterparts is less discernible, indicating that these models may not yet be suitable for dataset augmentation. Future research should focus on developing techniques to enhance the interpretability of these synthetic images for practical applications.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

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. *Advances in neural information processing systems*, 30, 2017.

Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Sharon Zhou. Conditional generation: Inputs, 2022. URL https://www.coursera.org/lecture/build-basic-generative-adversarial-networks-gans/conditional-generation-inputs-2OPrG.

A Appendix

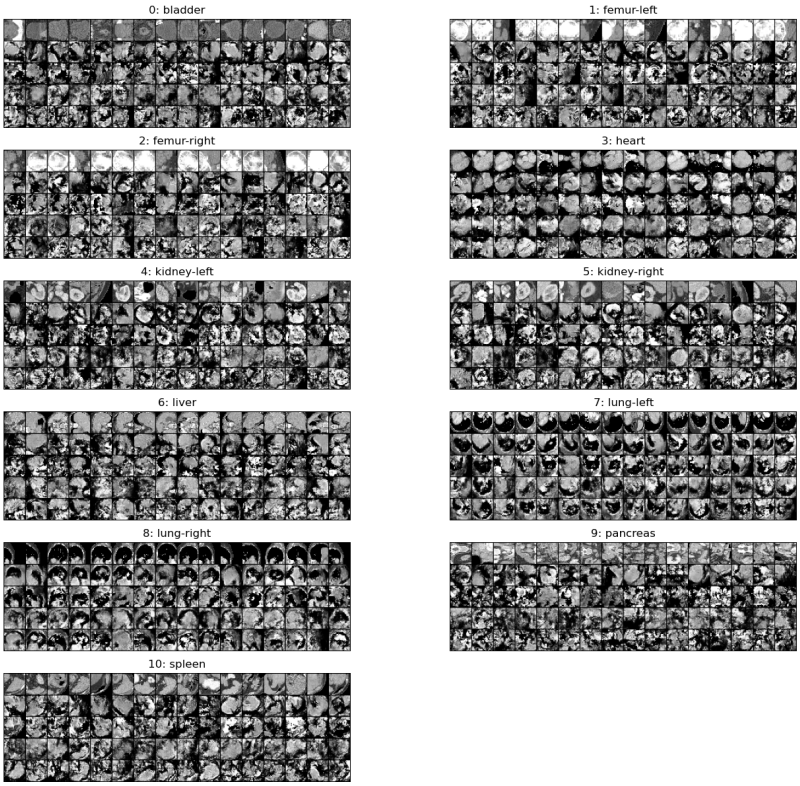

Figure 7: Generate results for OrganA data set

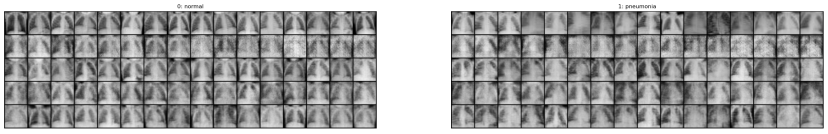

Figure 8: Generate results for Pneumonia data set

A.1 Analysis of FID score on OrganA Data set

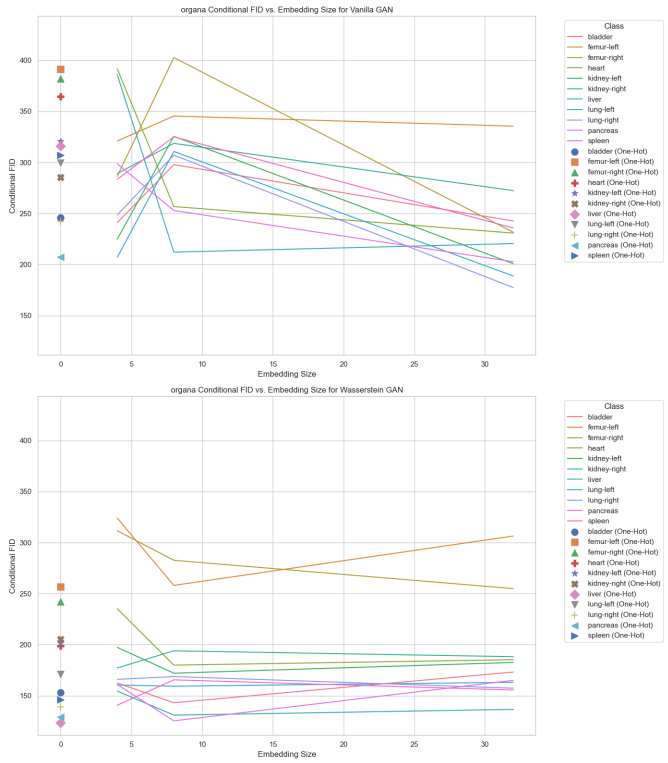

Figure 9: Conditional FID score vs. Embedding size for Organa data set

The Figure 9 shows that both cGANs exhibit a decrease in FID scores when the embedding size increases from 4 to 8. This trend is more pronounced in Vanilla cGAN, suggesting that more significant embeddings significantly enhance its ability to generate more accurate images of both normal and pneumonia-affected lungs.

The improvement in FID scores tends to plateau as the embedding size increases beyond 8. This indicates diminishing returns from increasing the embedding size to 32, particularly for Vanilla cGAN. The flattening of the decrease suggests that an 8-size embedding might be optimal for capturing the essential features of lung images in these models.

Wasserstein cWGAN shows a slight, consistent decrease in FID scores for the pneumonia class with larger embeddings, indicating its effectiveness in capturing the complex features of pneumonia. However, for the normal class, the FID score remains stable across embedding sizes, implying that Wasserstein cWGAN is less sensitive to changes in embedding size for generating images of healthy lungs.

One-hot encoding notably improves FID scores, especially in Vanilla cGAN, demonstrating its efficacy in enhancing image accuracy. In contrast, Wasserstein cWGAN does not show the same improvement with One-Hot encoding.

In summary, increasing the embedding size from 4 to 8 is crucial for both cGANs in learning relevant features, particularly for Vanilla cGAN. Beyond this point, the additional increase in embedding size shows limited benefits, especially for normal lung images, suggesting that an 8-size embedding might be sufficiently compelling for these scenarios.

According to the Figure 10, cWGAN with One-Hot encoding tends to perform better across multiple organ classes, indicating its effectiveness in capturing distinct organ features with explicit categorical representations. This suggests a general advantage in using discrete representations for complex organ imaging.

Sensitivity to embedding size varies notably among different organ classes. Some organs, like the bladder and spleen, show less sensitivity to changes in embedding size in the cWGAN, maintaining stable FID scores across different sizes. This implies that smaller embeddings may adequately capture the features of these organs.

Conversely, other organ classes, particularly the femur and liver, prefer larger embeddings or One-Hot encoding in the Wasserstein cWGAN. This trend highlights the complexity and diversity of features within these organs, requiring more significant embeddings or distinct categorical representations for effective image generation.

Vanilla cGAN shows varying degrees of sensitivity to embedding size changes across different organs, with some classes, like the heart and pancreas, exhibiting considerable sensitivity. In some cases, mid-range embeddings provide the best results, suggesting a nuanced relationship between embedding size and organ feature representation.

In summary, while the performance of cGAN models and the optimal embedding strategy varies across different organ classes, a general trend emerges where Wasserstein cWGAN with One-Hot encoding performs better in capturing organ features. The varying sensitivity to embedding sizes across organs underscores the need for tailored approaches in the cGAN model and embedding size selection for medical imagery generation.

A.2 Analysis of FID score on Pneumonia Data set

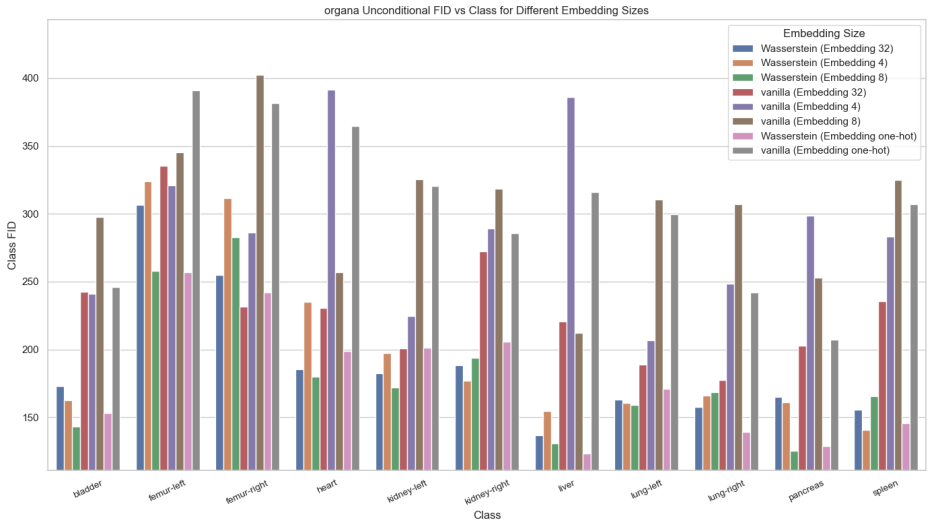

Figure 10: Conditional FID score vs. Classes for Organa data set

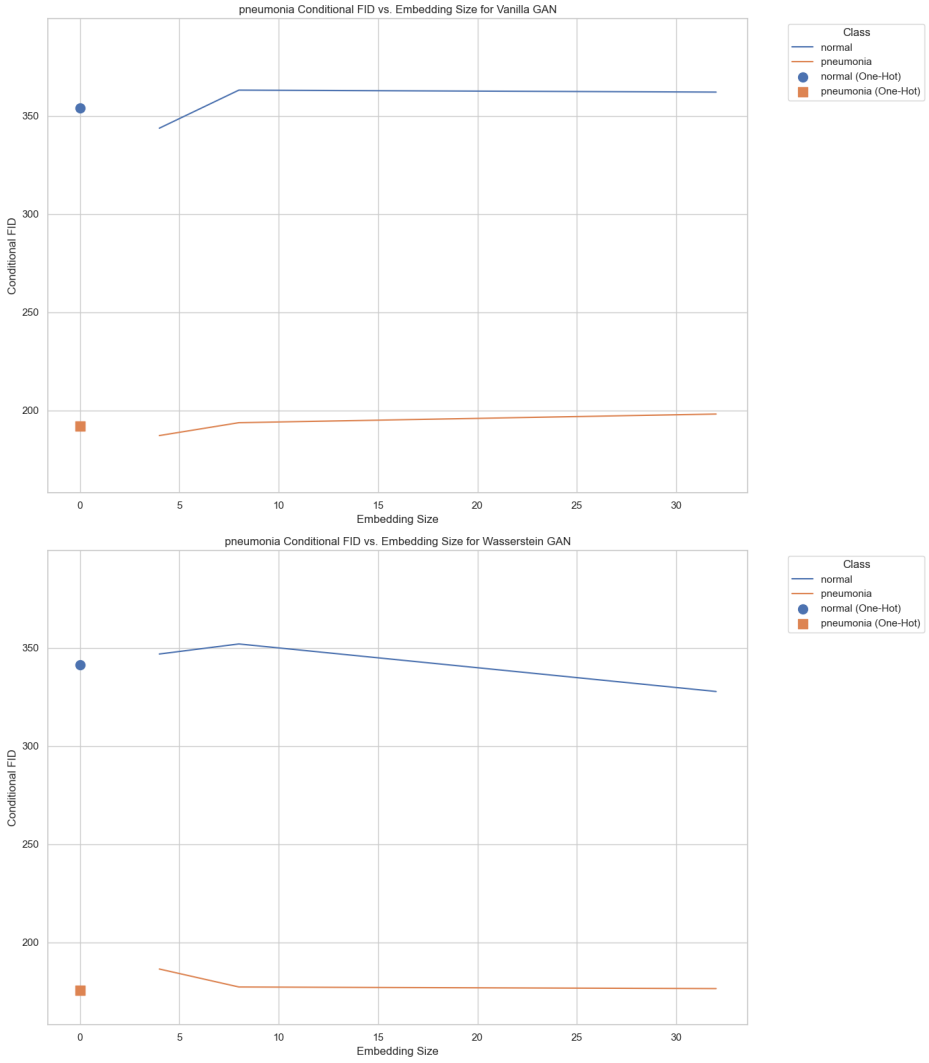

Figure 11: Conditional FID score vs. Classes for Pneumonia data set

Figure 11 shows a noticeable decrease in FID scores for both normal and pneumonia classes as the embedding size increases from 4 to 8. This indicates that a larger embedding size improves the Vanilla cGAN's ability to generate more accurate images, reducing the FID score. The decrease tends to flatten out as the embedding size exceeds 8, suggesting that increasing the embedding size to 32 offers diminishing returns in FID improvement.

The cWGAN shows a slight but consistent decrease in FID scores for the pneumonia class as the embedding size increases, indicating that larger embeddings can capture the complexity of pneumonia-related features more effectively. For the normal class, the FID score remains relatively stable across embedding sizes, suggesting that the Wasserstein cWGAN is less sensitive to embedding size for generating images of healthy lungs.

Therefore, both cGANs benefit from increased embedding size, but the effect is more pronounced in the vanilla cGAN. One-hot encoding demonstrates a clear advantage in vanilla cGAN, substantially improving the FID scores, while cWGAN does not show the same level of improvement.

Finally, the initial increase from embedding size 4 to 8 results in a significant decrease in FID for vanilla cGAN, suggesting that this range is critical for learning relevant features. Increasing the embedding size from 8 to 32 does not yield substantial improvements in FID scores for either cGAN,

especially for the typical class, indicating that embedding size 8 may be sufficient for capturing the necessary features for image generation in these cases.

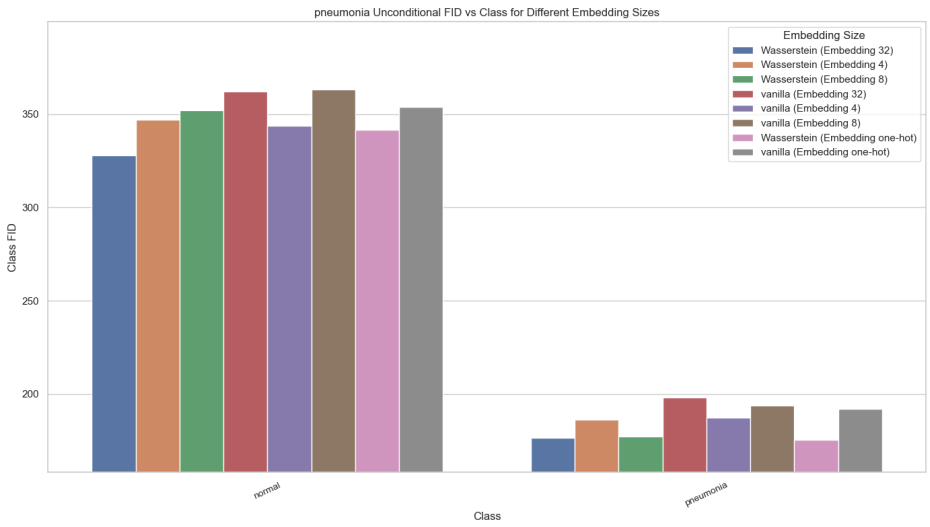

Figure 12: Conditional FID score vs. Embedding size for Pneumonia data set

The Figure 12 displays that FID scores across various embeddings are relatively close for the pneumonia class, particularly in the cWGAN, indicating lower sensitivity to embedding size changes. The vanilla cGAN shows more variation with embedding size in the pneumonia class, but the overall impact on the FID score is less marked than the normal class.

One-hot encoding in the cWGAN consistently yields the best FID scores for normal and pneumonia classes. This indicates the effectiveness of discrete, categorical representations in enhancing image synthesis quality, particularly for complex medical images.

In summary, the performance of cGANs in generating images of normal and pneumonia-affected lungs is significantly influenced by the choice of embedding size and encoding method. The Vanilla cGAN is more sensitive to changes in embedding size, while the cWGAN maintains a more consistent performance across different sizes. One-hot encoding, especially in the cWGAN, emerges as a beneficial approach for high-quality image generation, indicating its suitability for medical imaging tasks.