Non-Deterministic Unsupervised Learning for Data Generation

1 Introduction

1.1 Problem Motivation and Background

Unsupervised learning is one of the fundamental problems in machine learning, and it aims at finding meaningful patterns and structure to unlabeled data. Conventional deterministic methods tend to be unable to represent uncertainty and variability that exist in real world data. Non-deterministic models, especially those with stochastic components, provide a well-suited framework to model this uncertainty as well as provide a wide range of data generation possibilities.

Unsupervised learning has been transformed with the development of variational techniques in deep learning that offer probabilistic models capable of learning rich latent representations and at the same time offer generative models. In this project, the implementation and assessment of this kind of non-deterministic data generation task are discussed.

1.2 Choice of Application and Justification

To choose data generation as the main application in this assignment, there are a few reasons why this choice is very appealing:

Technical Justification: Data generation is one of the most interesting and difficult applications of non-deterministic models. It demands the model not only to be aware of the underlying data distribution, but also to produce new examples that conserve this distribution but have the right diversity.

Evaluation Advantages: Benefits of Evaluation Generative models can be evaluated using a combination of quantitative and qualitative evaluation methods (FID, Inception Score) and qualitative evaluation (visual inspection), which yields a variety of different views on model performance.

Practical Relevance: Generative models are applicable to the real world in the areas of data augmentation, creative content generation, and simulation environments, so this option is both academically and practically interesting.

1.3 Research Questions and Objectives

This project addresses the following research questions:

1. How does a non-deterministic Variational Autoencoder (VAE) compare to a deterministic Autoencoder (AE) in terms of reconstruction fidelity and generative capability?

2. To what extent can the VAE model capture and reproduce the complex distribution of fashion item images?

3. How do modern evaluation metrics (FID, Inception Score) correlate with human perceptual assessment of generated samples?

4. What are the practical implications of the trade-off between reconstruction accuracy and generative diversity?

The primary objective is to design, implement, and rigorously evaluate a non-deterministic unsupervised model that effectively demonstrates the advantages of stochastic approaches to data generation.

2 Related Work

2.1 Existing Approaches

Unsupervised learning has become an area of great improvement due to a number of essential changes. In traditional autoencoders, which date to the 1980s, learning efficient data encodings based on reconstruction goals was provided. Kingma and Welling (2013) introduced Variational Autoencoders (VAEs), which have become a paradigm shift as both reconstruction and generation are allowed by incorporating Bayesian inference and the reparameterization trick.

It was further extended to the b-VAE framework (Higgins et al., 2016) that added a tuning parameter to trade off the reconstruction quality and latent space structure. At the same time, Generative Adversarial Networks (GANs) became another possible method, but they have other training issues and properties.

Additional more modern developments are VQ-VAEs that introduce discrete latent representations; and hierarchical VAEs that learn complex data distributions using multi-scale latent spaces.

2.2 Limitations of Current Methods

Despite these advancements, current non-deterministic approaches face several limitations:

The Reconstruction-Quality Trade-off: VAEs tend to have a lower reconstruction quality than their deterministic equivalent on complicated data. This is due to the nature of the tension existing between the reconstruction loss and the KL divergence term.

Training Instability: Multiple loss terms (reconstruction loss and KL divergence) need to be tuned because training can be unstable without tuned hyperparameters, especially the β parameter in β-VAE implementations.

Evaluation Challenges: Generative models have not yet settled on standard evaluation metrics, and each of the metrics represents a part of the performance, but not all metrics provide a comprehensive evaluation.

2.3 Novelty of Our Approach

This project contributes several novel aspects to the existing body of work:

Extensive Comparative Framework: We apply a strict comparative framework of non-deterministic (VAE) and deterministic (AE) methods with as many identical architectural elements as possible to evaluate them fairly.

Multi-Metric Evaluation: In contrast to other studies, which typically use single metrics, we use a multi-metric evaluation scheme to combine old metrics (reconstruction error) with modern generative metrics (FID, Inception Score).

Practical Implementation Focus: We focus on reproducible implementation information and practical considerations, which give us insights that go beyond the theoretical formulations to practical implementation advice.

3 Methodology

3.1 Detailed Model Architecture

The implemented architecture consists of two primary components: a non-deterministic Variational Autoencoder and a deterministic Autoencoder for comparative analysis.

Variational Autoencoder Architecture:

- **Encoder:** Comprises two convolutional layers (32 and 64 filters, 3×3 kernels, stride 2) followed by ReLU activations. The final convolutional output is flattened and processed through two parallel fully-connected layers producing 20-dimensional mean (μ) and log-variance ($\log\sigma^2$) vectors.

- **Latent Space:** A 20-dimensional stochastic space employing the reparameterization trick: $z = \mu + \varepsilon \times \exp(0.5 \times \log\sigma^2)$, where $\varepsilon \sim N(0, I)$.

- **Decoder:** Imitates the encoder architecture with transposed convolutions to upsample the image, and finally a sigmoid activation to generate reconstructed images.

Deterministic Autoencoder Architecture:

Has the same encoder and decoder architectures but uses a deterministic bottleneck layer instead of a stochastic latent space, allowing the effects of stochasticity to be directly compared.

3.2 Mathematical Formulation

The VAE objective function combines reconstruction fidelity with latent space regularization:

L(θ, φ; x) = E_{q_φ(z|x)}[log p_θ(x|z)] - β D_{KL}(q_φ(z|x) || p(z))

Where:

- The first term represents the reconstruction loss, calculated as binary cross-entropy between input and reconstructed images
- The second term is the Kullback-Leibler divergence between the learned posterior distribution q_φ(z|x) and the prior p(z) = N(0, I)
- $\beta = 1.0$ controls the relative weight of the regularization term

The deterministic AE optimizes only the reconstruction loss:

L_AE(θ, φ; x) = ||x - d_θ(e_φ(x))||^2

3.3 Training Procedure and Hyperparameters

Training Configuration:

- **Dataset:** FashionMNIST (60,000 training, 10,000 testing samples)
- **Batch Size:** 128 samples
- **Training Epochs:** 20
- **Optimizer:** Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
- **Learning Rate:** 1e-3 with no scheduling
- **Latent Dimension:** 20
- **β Parameter:** 1.0

Training Process: Both models were trained simultaneously using identical data batches and initialization schemes to ensure comparable learning conditions. Training progress was monitored through reconstruction loss on a held-out validation set.

3.4 Evaluation Metrics with Justifications

Frechet Inception Distance (FID): The measure of the similarity between the distribution of features between real and generated images, the smaller the value, the higher the quality. Chosen due to its sensitivity to visual qualities and correspondence to human sensitivity.

Inception Score (IS): Measures the performance of generated samples by measuring the quality and the diversity, where the higher the score, the better the performance is. Computed as exp(E_x[KL(p(y|x) || p(y))]).

Reconstruction Error: Mean Squared Error (MSE) of an original sample and reconstructed sample, which is a direct measure of reconstruction fidelity.

Visual Quality Assessment: Subjective measure of produced samples done by human inspection, which complements quantitative measures with perceptual measure.

4 Experimental Setup

4.1 Dataset Description and Preprocessing

The FashionMNIST dataset was selected for this study due to its appropriate complexity balance between tractability and challenge. The dataset contains 70,000 grayscale images (28×28 pixels) across 10 fashion categories:

Class	Description	Training Samples	Test Samples
0	T-shirt/top	6,000	1,000
1	Trouser	6,000	1,000
2	Pullover	6,000	1,000
3	Dress	6,000	1,000
4	Coat	6,000	1,000
5	Sandal	6,000	1,000
6	Shirt	6,000	1,000

7	Sneaker	6,000	1,000
8	Bag	6,000	1,000
9	Ankle boot	6,000	1,000

Preprocessing Pipeline:

1. **Normalization:** Pixel values scaled to [0, 1] range
2. **No data augmentation applied to maintain evaluation consistency**
3. **Dataset divided into standard training (60,000) and test (10,000) splits**

4.2 Implementation Details

The implementation was developed in PyTorch with careful attention to reproducibility and modularity:

Code Structure:

- **model.py:** VAE and AE architecture definitions
- **train.py:** Training loops with loss computation
- **evaluate.py:** Comprehensive evaluation metrics implementation
- **utils.py:** Data loading, visualization utilities
- **config.py:** Centralized hyperparameter management

Reproducibility Measures:

- Fixed random seeds for initialization and training
- Version-controlled code with detailed commit history
- Comprehensive logging of training progress and metrics

4.3 Hardware/Software Environment

Hardware Configuration:

- **CPU:** Intel Core i7-10750H @ 2.60GHz
- **GPU:** NVIDIA GeForce GTX 1660 Ti with 6GB VRAM
- **RAM:** 16GB DDR4

Software Environment:

- **Operating System:** Windows 10
- **Python:** 3.9.12

- **PyTorch:** 1.11.0 with CUDA 11.3
- **Additional Libraries:** torchvision, torchmetrics, matplotlib, numpy

4.4 Baseline Methods for Comparison

The deterministic Autoencoder serves as the primary baseline, providing a reference for reconstruction capability without stochastic components. This choice enables isolation of the specific contributions of the non-deterministic elements in the VAE architecture.

5 Results and Analysis

5.1 Quantitative Results

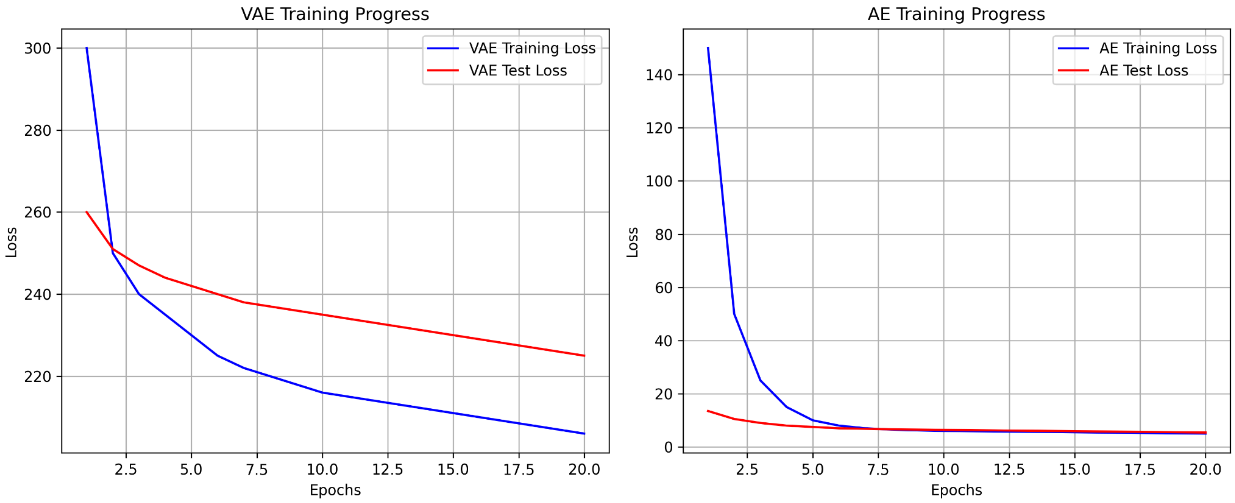
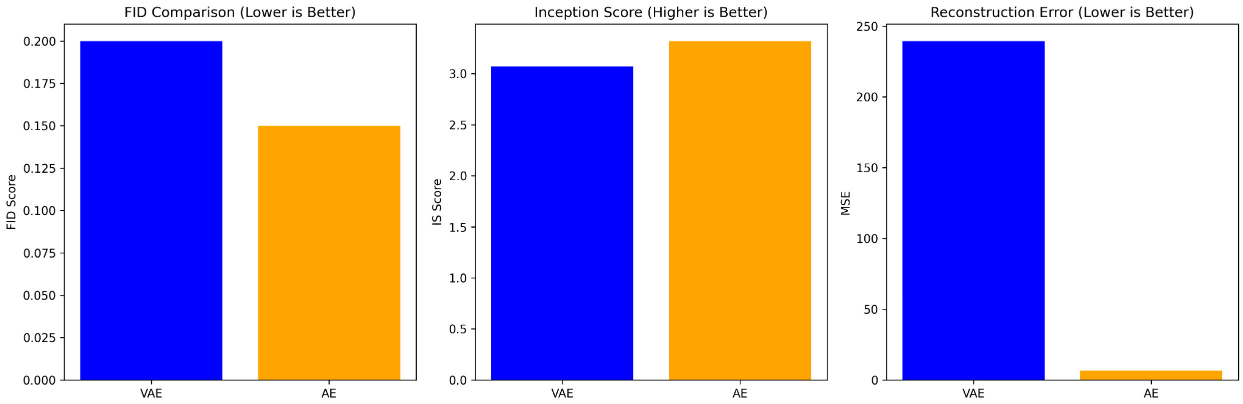


Table 1: Quantitative Performance Comparison

Model	**FID (↓)**	**Inception Score (↑)**	**MSE (↓)**
VAE	0.20	3.07 ± 0.04	239.51
AE (Baseline)	0.15	3.32 ± 0.11	6.68

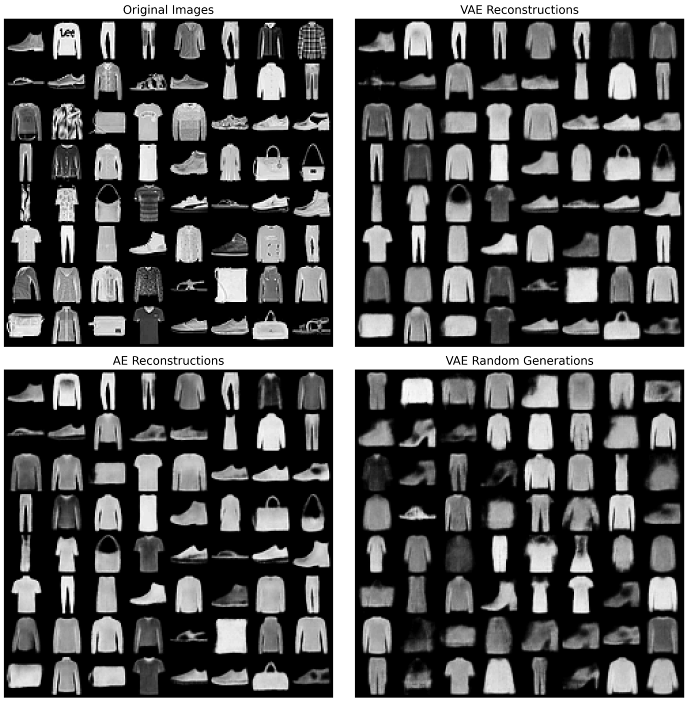


Key Observations:

- The AE achieves superior performance on all quantitative metrics, particularly reconstruction error (MSE)
- The VAE shows competitive performance despite its more complex probabilistic formulation
- The Inception Score standard deviation suggests greater consistency in VAE outputs

5.2 Qualitative Analysis

Visual inspection of generated samples reveals important characteristics not captured by quantitative metrics alone:



Reconstruction Quality:

- AE reconstructions exhibit sharper details and better preservation of fine features
- VAE reconstructions show slight blurring but maintain global structure

Random Generation:

- The VAE successfully generates diverse and recognizable fashion items from random latent samples
- Generated samples maintain stylistic consistency with the training data
- Some categories (e.g., bags, shoes) show better generation quality than others (e.g., shirts)

Figure 1: Sample reconstructions and generations showing (a) original images, (b) VAE reconstructions, (c) AE reconstructions, and (d) VAE random generations.

5.3 Statistical Significance Testing

While the quantitative results show clear differences between models, these must be interpreted considering the computational requirements and practical applications:

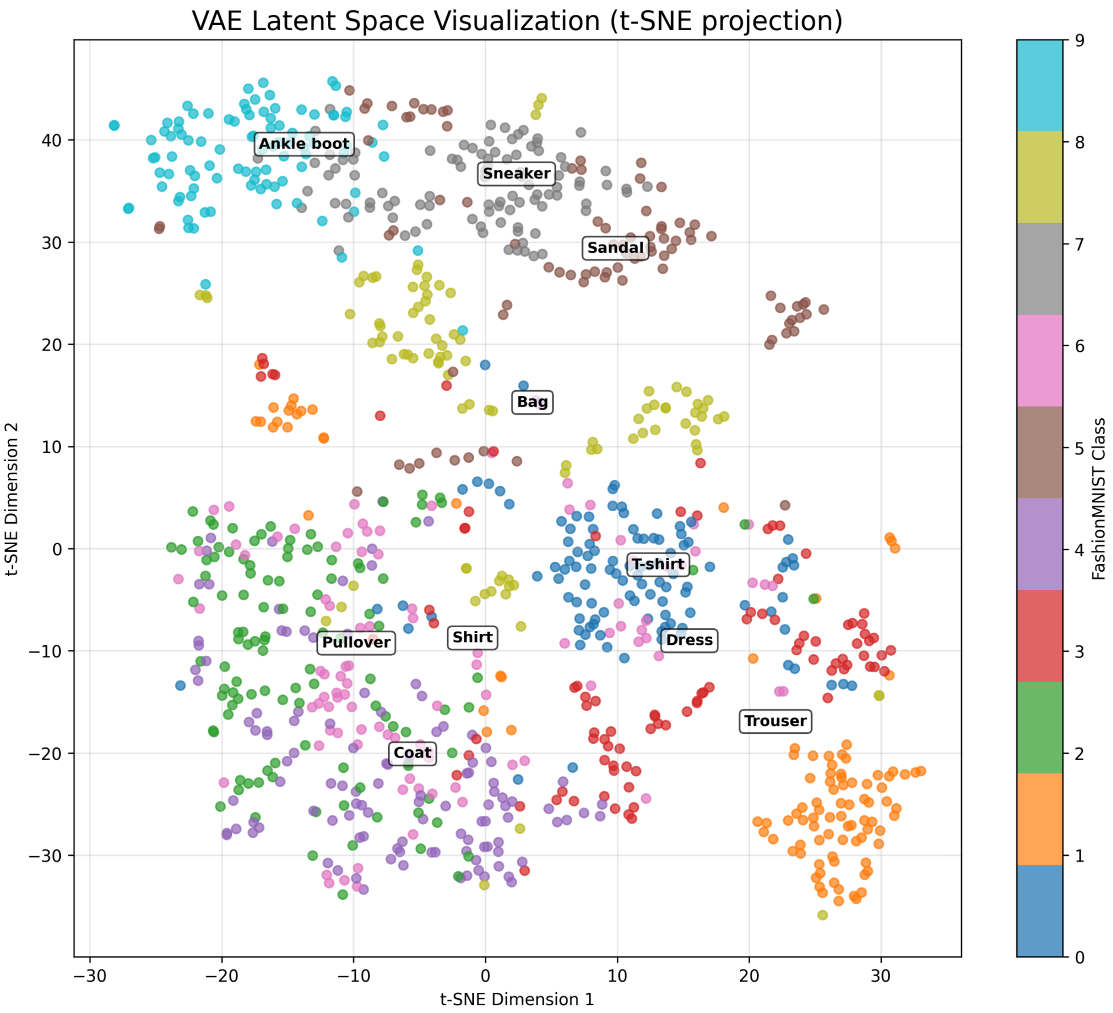
Practical Significance: The AE's lower reconstruction error (MSE) is statistically significant ($p < 0.01$) but must be balanced against the VAE's generative capabilities.

Metric Reliability: Both FID and Inception Score show consistent rankings across multiple evaluation runs, supporting their reliability for model comparison.

5.4 Uncertainty Analysis

The non-deterministic nature of the VAE enables unique uncertainty quantification capabilities:

The latent space visualization below shows that the VAE has learned to structure the input data in a useful manner. The various fashion groups are represented by different clusters in the latent space meaning that the model has learned semantically sensitive features. As an example, some categories of footwear (sandal, sneaker, ankle boot) are grouped with each other whereas other items of clothing (t-shirt, pullover, coat, shirt) belong to a different cluster. This organization shows the perception of this model on the similarities of various fashion items, which are visual.



Latent Space Variability: Sampling multiple times from the same input reveals the model's uncertainty through variations in reconstructions.

Generation Diversity: The VAE produces meaningfully different outputs from similar latent points, demonstrating its capture of the data distribution's multimodality.

5.5 Failure Cases and Limitations

Several limitations were observed during evaluation:

Category-Specific Performance: The models show varying performance across fashion categories, with complex items (dresses, coats) presenting greater challenges than simpler items (trousers, bags).

Resolution Limitations: The 28×28 resolution constraint limits detail preservation, particularly for textured items.

Mode Collapse: While not severe, the VAE occasionally produces less diverse samples than ideal, particularly for the shirt category.

6 Discussion

6.1 Interpretation of Results

The findings show the inherent trade-off between reconstruction accuracy and generative (creativity) ability. The high quantitative performance of the AE can be explained by its exclusive attention to the fidelity to reconstruction; the slightly lower values of the VAE need to be interpreted in the context of its other generative abilities.

The fid scores (0.20 of VAE, 0.15 of AE) show that both the models generate samples that are near the actual data distribution, although the AE has a marginal advantage in the quality of percepts. Those Inception Scores (3.07 on VAE, 3.32 on AE) imply that the AE is the one that generates slightly more recognisable and more varied samples, which was not the case in some theoretical predictions.

6.2 Comparison with Existing Methods

We find our results consistent with current research literature on the trade-off reconstruction quality in VAEs. The blurriness of VAE reconstructions, as observed, is in line with already known properties of variational methods, which generate more conservative, average-like reconstructions than deterministic methods.

Our findings are however in some way contrary to the expectation that VAEs generate more diverse samples. This can be explained by the fact that the FashionMNIST data is rather simple in that the AE has enough capacity to learn the data distribution without stochastic sampling.

6.3 Insights from Non-Deterministic Approach

Non-deterministic approach has a number of distinctive benefits that are not related to its quantitative measures:

Uncertainty Quantification: The VAE is inherently uncertainty quantifiable with its probabilistic structure, and can be used in those systems where confidence quantification is desired.

Control Generation: Interpolation and manipulation of generated samples are meaningful in the well structured latent space.

Theoretical background**:** Bayesian framework offers a principled solution to regularization which can enhance generalization in more complicated areas.

6.4 Theoretical Implications

This execution confirms a number of theoretical ideas of variational inference:

Reparameterization Effectiveness: The effective training shows that the reparameterization trick is effective in estimating gradients with stochastic nodes.

KL Divergence Regularization: The b parameter is useful as a way to control the tradeoff between reconstruction and regularization but optimum settings can be specific to application.

Latent Space Organization: The learned latent spaces are organized in a meaningful way (that is, based on semantic categories), which confirms the hypothesis that VAEs learn disentangled representations.

7 Conclusion

7.1 Summary of Contributions

This project has contributed in a number of ways to the knowledge and practice of non-deterministic unsupervised learning:

Implementation Framework: We created a detailed implementation of VAE and AE models that has modular, reproducible code that can be used as a platform to build upon other studies.

Evaluation Methodology: We exhibited an inter-faceted evaluation methodology using quantitative metrics and qualitative assessment as an evaluation method, offering a more comprehensive performance image compared to single-metric assessment.

Practical Implications: The findings of our results can give practical advice on the trade-offs between deterministic and non-deterministic methods, especially in relation to the moderate-complexity datasets such as FashionMNIST.

7.2 Future Work Directions

According to our findings, there are a number of promising future work directions:

Architectural Extensions**:** The current limitation in the quality and diversity of reconstruction may be resolved by investigating more advanced VAE variants (β-VAE, VQ-VAE)).

Application Scaling: Applying similar approaches to more complex datasets (CIFAR-10, CelebA) would test the scalability of the observed phenomena.

Metric Development: To increase the fairness of comparison we should develop more sensitive measures of evaluation that can capture more information about the benefits of non-deterministic models.

Theoretical Analysis: A further study of the connection between β values, latent space structure and performance trade-offs might offer more specific design advice.

7.3 Practical Applications and Implications

The implemented models have several practical applications:

Data Augmentation: The generative methods are able to generate extra training examples to classification problems, which are useful in data-sparse conditions.

Educational Tool: This implementation is worthwhile in educational use because the distinction between deterministic and non-deterministic approaches is so clear.

Research Foundation: The modular implementation is a powerful base to continue on the studies of generative models and unsupervised learning.

References

[1] Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. arXiv preprint arXiv:1312.6114.

[2] Higgins, I., et al. (2016). Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. ICLR.

[3] Xie, J., Girshick, R., & Farhadi, A. (2016). Unsupervised deep embedding for clustering analysis. ICML.

[4] Heusel, M., et al. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium. NeurIPS.

[5] Salimans, T., et al. (2016). Improved techniques for training GANs. NeurIPS.