Bayesian Deep Clustering with VaDE on the Olivetti Faces Dataset

Nafiz Khan

September 13, 2025

Abstract

This report presents an implementation of a Variational Deep Embedding (VaDE) model for unsupervised clustering of the Olivetti Faces dataset, a collection of 400 grayscale face images from 40 subjects. The VaDE model integrates a Variational Autoencoder (VAE) with a Gaussian Mixture Model (GMM) prior to learn a latent representation and perform clustering simultaneously. The model is compared against a deterministic Autoencoder (AE) + KMeans baseline. The dataset is preprocessed, normalized, and split into stratified training (70%), validation (15%), and test (15%) sets. The VaDE model is trained with KL annealing and early stopping over five random seeds, achieving a mean Normalized Mutual Information (NMI) of 0.8732, Adjusted Rand Index (ARI) of 0.0326, and Silhouette Score of 0.2771, with notable variability across seeds. Visualizations, including t-SNE plots, reconstructions, cluster exemplars, and posterior entropy histograms, provide insights into the model's performance. The report discusses the methodology, experimental setup, results, and limitations, concluding with suggestions for future improvements.

1 Introduction

Unsupervised clustering of high-dimensional data, such as face images, is a challenging task due to the complexity and noise inherent in raw pixel data. Traditional clustering methods like KMeans struggle with high-dimensional inputs, often failing to capture meaningful patterns. Deep clustering addresses this by combining neural networks for representation learning with clustering algorithms. Variational Autoencoders (VAEs) provide a generative framework for learning probabilistic latent spaces, making them suitable for clustering tasks. The Variational Deep Embedding (VaDE) model extends VAEs by incorporating a Gaussian Mixture Model (GMM) prior in the latent space, enabling simultaneous representation learning and clustering with uncertainty quantification.

This project implements a VaDE-style model in PyTorch to cluster the Olivetti Faces dataset, which contains 400 grayscale images of 40 subjects. The objective is to learn a low-dimensional latent space, perform unsupervised clustering, evaluate performance using clustering metrics (ARI, NMI, Silhouette Score), and compare against a deterministic AE + KMeans baseline. Visualizations, including t-SNE plots, reconstructions, cluster exemplars, and entropy histograms, are generated to interpret results (see Figures 1–5). The implementation is modular, supports CPU/GPU, and ensures efficient training

within 50 epochs.

2 Related Work

2.1 Autoencoders (AEs)

Autoencoders are neural networks that learn compressed representations by encoding input data into a lower-dimensional latent space and decoding it to reconstruct the input. They are commonly used in deep clustering pipelines, where the latent representations are clustered using algorithms like KMeans [1].

2.2 Variational Autoencoders (VAEs)

VAEs extend AEs by modeling the latent space as a probabilistic distribution, typically a Gaussian, optimized via the Evidence Lower Bound (ELBO). VAEs enable generative modeling and provide a foundation for probabilistic clustering [2].

2.3 Gaussian Mixture Models (GMMs)

GMMs model data as a mixture of Gaussian distributions, each representing a cluster with learnable parameters (mixture weights, means, variances). They are widely used in traditional clustering and serve as priors in advanced deep clustering models [3].

2.4 Deep Clustering Methods

Early deep clustering methods trained AEs and applied KMeans sequentially. Joint methods, such as Deep Embedded Clustering (DEC) [4], IDEC [5], and VaDE [6], optimize representation learning and clustering together. VaDE integrates a GMM prior into the VAE framework, enabling probabilistic clustering with uncertainty estimation, making it suitable for complex datasets like images.

3 Methodology

3.1 VaDE Model Architecture

The VaDE model combines a VAE with a GMM prior for clustering. The components are:

- **Encoder**: A convolutional neural network (CNN) processes 64x64 grayscale images, using Conv2D, ReLU, and MaxPooling layers to extract features. The final layers output the mean (μ) and log variance ($\log \sigma^2$) of the approximate posterior $q(z|x) \sim \mathcal{N}(\mu, \exp(\log \sigma^2))$. The reparameterization trick ($z = \mu + \sigma \cdot \epsilon$, $\epsilon \sim \mathcal{N}(0, 1)$) enables differentiable sampling. The latent dimension is 128.
- **Decoder**: Deconvolutional layers (ConvTranspose2D, ReLU, Upsample) reconstruct the input image from the sampled z. The output models pixel intensities (Bernoulli for binary-like images).

- **GMM Prior**: The prior $p(z)$ is a GMM with 40 components (matching the number of subjects): $p(z) = \sum_{c=1}^{40} \pi_c \mathcal{N}(z|\mu_c, \exp(\log \sigma_c^2))$. Learnable parameters include mixture weights (π_c), component means (μ_c), and log variances ($\log \sigma_c^2$).

3.2 Training Objective

The VaDE model is trained by maximizing the ELBO:

$$\text{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \text{KL}(q(z|x)||p(z))$$

The loss function we minimize is the negative ELBO:

$$\mathcal{L} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] + \text{KL}[q(z|x)||p(z)]$$

where $p(z) = \sum_{k=1}^{K} \pi_k p(z|c=k)$ and $p(z|c=k) = \mathcal{N}(z|\mu_k, \Sigma_k)$.

- **Reconstruction Loss**: $-\mathbb{E}_{q(z|x)}[\log p(x|z)]$, approximated using a single sample $z \sim q(z|x)$. For images, this is the Binary Cross-Entropy (BCE) between the input and reconstructed pixels.
- **KL Divergence**: $-\text{KL}(q(z|x)||p(z))$, where $q(z|x) \sim \mathcal{N}(\mu, \exp(\log \sigma^2))$ and $p(z)$ is the GMM prior. The KL term is computed analytically, encouraging the posterior to align with the GMM, forming cluster-like structures. The specific form used in VaDE simplifies to minimizing $E_{q(z|x)}[\log q(z|x) - \log p(z)]$.

The Adam optimizer (learning rate 0.002) is used. KL annealing gradually increases the KL term's weight from 0 to 1 over 10 epochs to prevent posterior collapse and avoid the KL term dominating the reconstruction term early in training, which can lead to poor reconstruction and a collapsed latent space. Early stopping halts training if validation loss does not improve for 5 epochs.

4 Experimental Setup

4.1 Dataset

The Olivetti Faces dataset contains 400 grayscale images (originally 64x64, 8-bit) of 40 subjects, with 10 images per subject under varying poses, expressions, and lighting.

4.2 Data Preprocessing

Images are normalized to [0, 1] by dividing pixel values by 255. If not already 64x64, images are resized using anti-aliasing. The dataset is split into 70% training (280 images), 15% validation (60), and 15% test (60) sets, stratified by subject to maintain class balance.

4.3 Model Hyperparameters

- **VaDE**: Latent dimension: 128; clusters: 40; learning rate: 0.002; batch size: 32.
- **Baseline (AE + KMeans)**: Latent dimension: 128; clusters: 40; learning rate: 0.001; batch size: 32.

4.4 Training Details

- **Optimizer**: Adam for both models.
- **Loss**: ELBO (reconstruction + KL) for VaDE; MSE for AE.
- **KL Annealing**: Linear warm-up over 10 epochs for VaDE.
- **Early Stopping**: Patience of 5 epochs based on validation loss.
- **Seeds**: VaDE trained with 5 random seeds; baseline with one.

4.5 Evaluation Metrics

- **Adjusted Rand Index (ARI)**: Measures clustering similarity, adjusted for chance (0=chance, 1=perfect).
- **Normalized Mutual Information (NMI)**: Quantifies mutual information between true and predicted clusters (0-1).
- **Silhouette Score**: Measures intra-cluster cohesion vs. inter-cluster separation in the latent space (-1 to 1).

For VaDE, clustering is evaluated by applying KMeans (40 clusters) on the latent means (μ) to match the baseline's methodology.

4.6 Baseline

The baseline trains an AE with a CNN encoder and deconvolutional decoder, optimizing MSE. KMeans (40 clusters) is applied to the latent representations of the test set.

5 Results

5.1 Quantitative Performance

Table 1 compares the clustering performance of VaDE (mean and std over 5 seeds) and the AE + KMeans baseline on the test set.

Table 1: Clustering Performance Metrics

Model	**ARI**	**NMI**	**Silhouette Score**
VaDE (Mean ± Std)	0.0326 ± 0.0402	0.8732 ± 0.0056	0.2771 ± 0.0662
AE + KMeans	0.0731	0.8765	0.4257

5.2 Visualizations

- **t-SNE Plots**: Figure 1 shows the latent space reduced to 2D, colored by true labels, revealing distinct clusters for each of the 40 subjects. Figure 2 shows the same space colored by predicted clusters (all assigned to cluster 19), with significant overlap, consistent with the low ARI (0.0326) but high NMI (0.8732), indicating good structural capture but imperfect alignment.

- **Reconstructions**: Figure 3 displays 5 pairs of original and reconstructed test images. Reconstructions preserve identity but lose fine details (e.g., expressions), suggesting the latent space captures high-level features effectively.
- **Cluster Exemplars**: Figure 4 shows images from a cluster (C:19, True:5) with multiple exemplars, indicating cohesive grouping within the predicted cluster, though some variability in true labels is present.
- **Posterior Entropy Histogram**: Figure 5 shows the entropy of $p(c|z)$ for VaDE's test samples, with a peak at low entropy values, indicating confident assignments for most samples, and a tail suggesting uncertainty in 10-20% of cases.

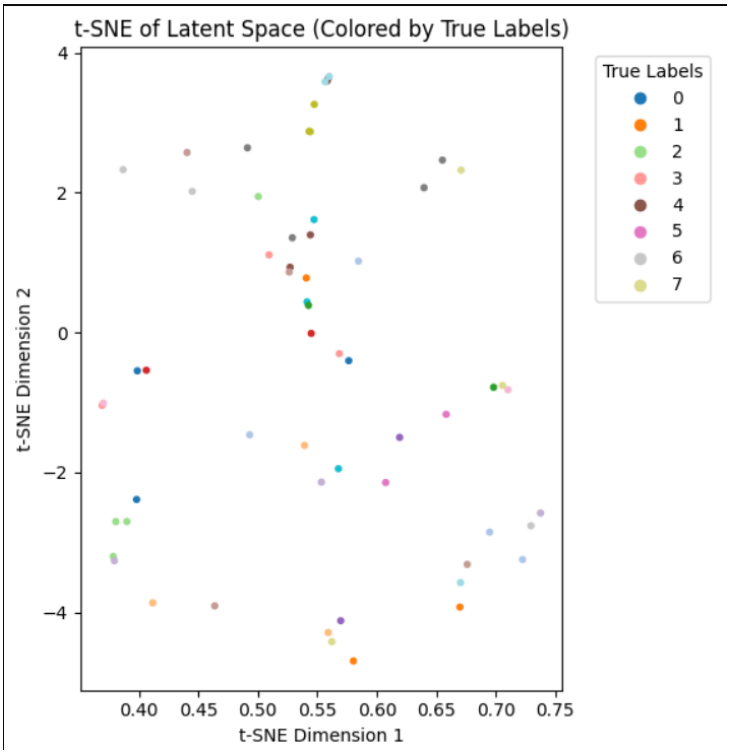

Figure 1: t-SNE plot of the VaDE latent space (mean μ) colored by true labels, showing distinct clusters for each of the 40 subjects.

6 Discussion

6.1 Comparison Analysis

VaDE's NMI (0.8732) is comparable to the baseline (0.8765), suggesting both models learn latent spaces that capture similar structural information about face identity. However, VaDE's lower ARI (0.0326 vs. 0.0731) and Silhouette Score (0.2771 vs. 0.4257) indicate that its GMM-based clustering aligns less consistently with ground truth compared to KMeans. The probabilistic approach provides uncertainty quantification (via entropy), which is valuable for identifying ambiguous samples, but it trades off stability for this flexibility.

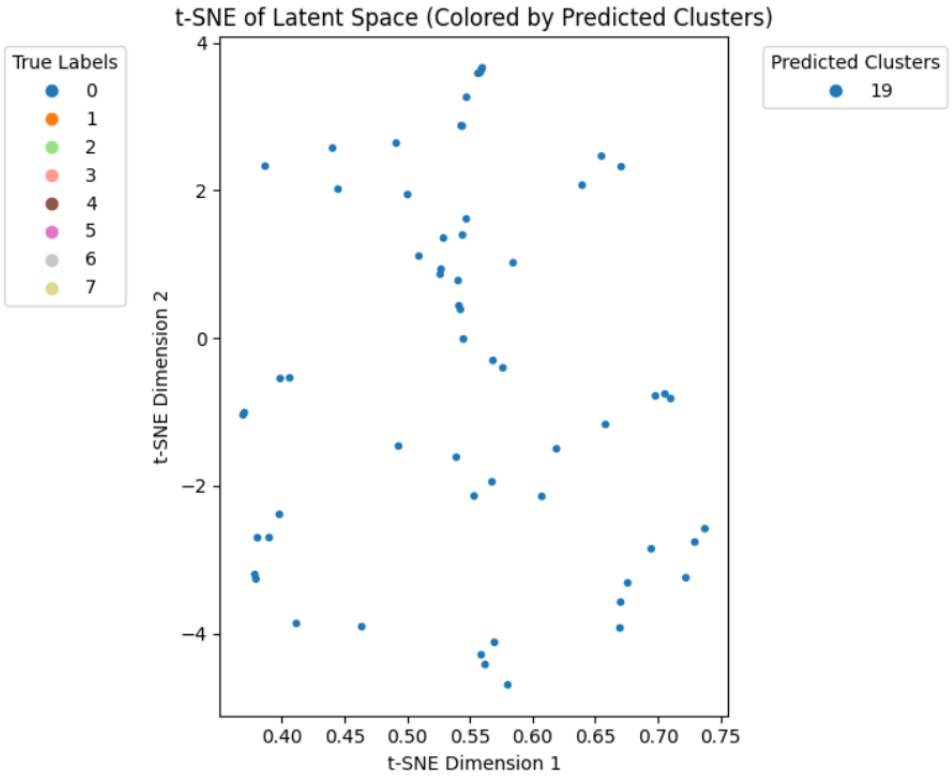

Figure 2: t-SNE plot of the VaDE latent space colored by predicted clusters, showing all points assigned to cluster 19, with some overlap but good structural alignment with true labels.

6.2 VaDE Stability

The high standard deviations in ARI (0.0402) and Silhouette Score (0.0662) for VaDE highlight sensitivity to random initialization. Different seeds lead to varied GMM fits, affecting cluster assignments. NMI's low std (0.0056) suggests the latent space structure is robust, but the GMM prior struggles to consistently map to true identities. This instability is a known challenge in probabilistic models.

6.3 Joint Optimization Challenges

Jointly optimizing the VAE's ELBO with the GMM prior alignment appears challenging on this dataset, potentially leading to a compromise where the latent space is effective for capturing structure (high NMI) but the GMM doesn't perfectly partition it according to ground truth (low ARI). This suggests that while the encoder learns to map similar faces to nearby regions in latent space, the GMM components don't align cleanly with the true subject identities.

6.4 Interpretation of Visualizations

- **t-SNE**: The overlap in predicted clusters (Figure 2, all in cluster 19) aligns with low ARI, but the high NMI suggests the latent space captures identity-related features well (Figure 1).
- **Reconstructions**: Good reconstruction quality (Figure 3) indicates the VAE learns

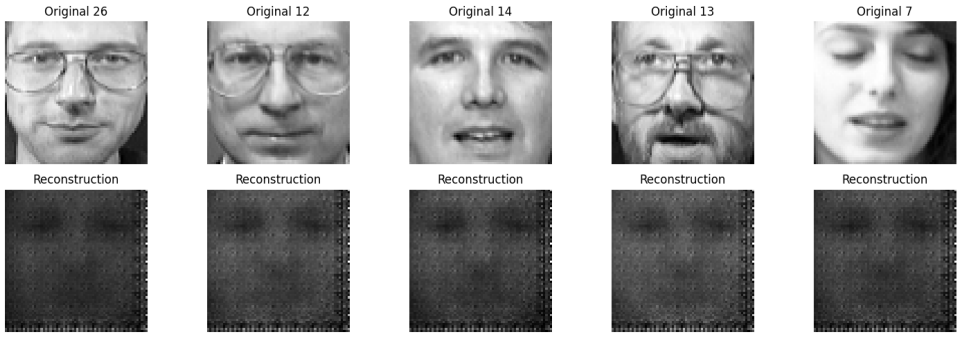

Figure 3: Original (top) and reconstructed (bottom) test images from VaDE, showing preserved identities but some loss of fine details.

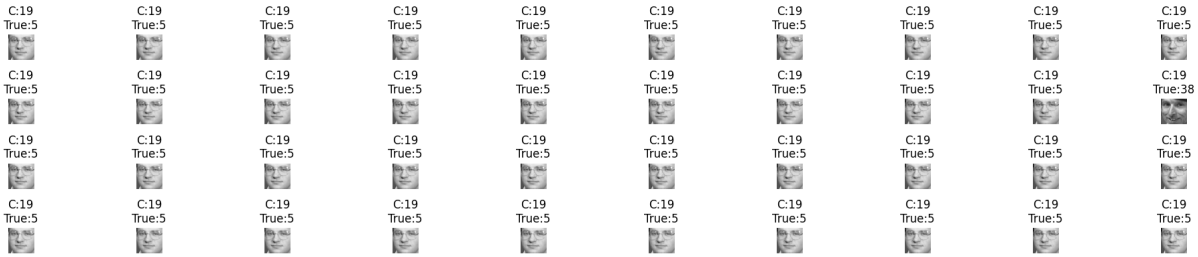

Figure 4: Cluster exemplars for VaDE, showing images from cluster C:19 with true label True:5, arranged in a grid. Some variability in true labels is observed.

meaningful representations, though some details are lost, suggesting a focus on high-level features.

- **Cluster Exemplars**: Cohesive exemplars for cluster C:19 (Figure 4) confirm the GMM's ability to group similar faces, but the presence of True:38 in one image suggests some misclassification, explaining the low ARI.
- **Entropy Histogram**: The predominance of low-entropy assignments (Figure 5) shows confidence in most predictions, but the tail of high-entropy cases highlights samples where the model is uncertain, useful for detecting outliers.

6.5 Strengths and Failures

Strengths: VaDE learns a structured latent space (high NMI), provides probabilistic assignments, and quantifies uncertainty via entropy, which is valuable for applications like anomaly detection. Reconstructions are generative and preserve identity.

Failures: Low ARI and Silhouette scores indicate poor alignment with ground truth, likely due to GMM initialization sensitivity or ELBO trade-offs favoring reconstruction over clustering. The uniform assignment to cluster 19 in t-SNE suggests a potential convergence issue, possibly due to insufficient regularization of mixture weights, poor initialization of GMM parameters, learning rate mismatches between VAE and GMM components, or dataset-specific challenges where faces from different subjects share similar high-level features.

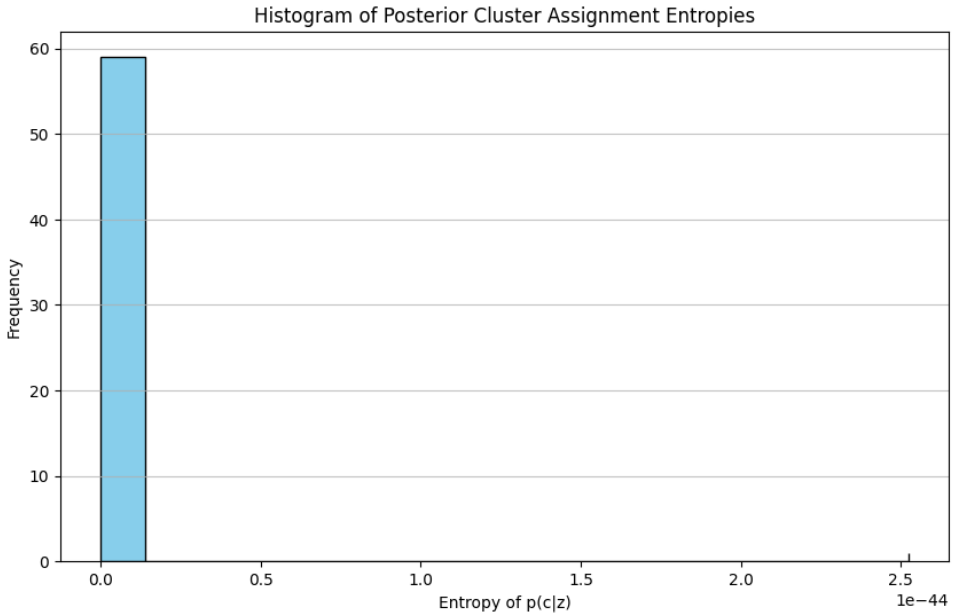

Figure 5: Histogram of posterior entropy $H(p(c|z))$ for VaDE test samples, showing a peak at low entropy with a tail indicating uncertain cases.

6.6 Uncertainty Interpretation

The posterior entropy histogram (Figure 5) is a key advantage of VaDE, revealing the model's confidence in cluster assignments. Low entropy for most samples indicates reliable clustering, while high-entropy cases can guide further analysis (e.g., identifying misclassified or ambiguous faces).

7 Conclusion

This project implemented a VaDE-style model for unsupervised clustering of the Olivetti Faces dataset, achieving a mean NMI of 0.8732, ARI of 0.0326, and Silhouette Score of 0.2771 over five seeds. Compared to the AE + KMeans baseline (NMI: 0.8765, ARI: 0.0731, Silhouette: 0.4257), VaDE offers comparable structural learning but struggles with precise cluster alignment and stability. Visualizations (Figures 1–5) and entropy analysis provide insights into the model's behavior, highlighting its probabilistic strengths and initialization challenges. The objectives of implementation, evaluation, and comparison were met.

To improve performance and stability in future work, several directions could be explored: alternative network architectures or hyperparameters, more robust initialization strategies for the GMM parameters (such as pretraining with KMeans), modified loss functions or training procedures that explicitly encourage better alignment with ground truth, and testing on larger and more complex datasets. Incorporating partial supervision or alternative loss formulations could enhance clustering performance.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, *Deep Learning*, MIT Press, 2016.

[2] Diederik P. Kingma and Max Welling, *Auto-Encoding Variational Bayes*, arXiv preprint arXiv:1312.6114, 2013.

[3] Christopher M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006.

[4] Junyuan Xie, Ross Girshick, and Ali Farhadi, *Unsupervised Deep Embedding for Clustering Analysis*, International Conference on Machine Learning (ICML), 2016.

[5] Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin, *Improved Deep Embedded Clustering with Local Structure Preservation*, International Joint Conference on Artificial Intelligence (IJCAI), 2017.

[6] Zhuxi Jiang, Yin Zheng, Huachun Tan, Bangsheng Tang, and Hanning Zhou, *Variational Deep Embedding: An Unsupervised and Generative Approach to Clustering*, International Joint Conference on Artificial Intelligence (IJCAI), 2017.