Hierarchical Variational Autoencoder for Low-Resolution Face Generation

Enock Onkarabile Buys

[1] Student Number:219013044
[2] Academy of Computer Science and software Engineering, University of Johannesburg

Abstract.

Generative models for face synthesis, such as StyleGAN and diffusion models, have predominantly focused on achieving high-resolution outputs, a process that demands substantial computational resources. This research addresses a gap in the literature by investigating the performance of a Hierarchical Variational Autoencoder (HVAE) for low-resolution (32×32 pixel) face generation, a relevant task for resource-constrained environments. The model implements a four-layer encoder-decoder architecture with progressive dimensional reduction. Trained on a subset of 5,000 images from the CelebA dataset over 600 epochs, the model demonstrates stable convergence. Quantitative evaluation yields a final reconstruction loss of 0.0726 and an approximate Fréchet Inception Distance (FID) score of 170.23. While the high FID and characteristic blurriness confirm known limitations of VAE architectures, the results demonstrate the feasibility of learning basic facial representations under severe computational limits. The study provides empirical benchmarks and a critical analysis of the trade-offs between pixel-wise accuracy and perceptual quality in low-resolution VAE-based generation.

Keywords: Variational Autoencoder, Low-Resolution, Face Generation, Computational Constraints, Generative Models.

1 Introduction

1.1 Background

Generative modeling has emerged as a critical area in computer vision, with applications spanning face synthesis, data augmentation, and privacy-preserving data generation. While significant advances have been made in high-resolution image generation through architectures like StyleGAN (Karras et al., 2019) and diffusion models (Ho et al., 2020), there exists a substantial gap in understanding generative model performance under computational constraints. Low-resolution face generation at 32×32 pixels presents unique challenges: preserving essential facial features while working within limited representational capacity.

Traditional Variational Autoencoders (VAEs) provide stable training through their probabilistic framework but suffer from reconstruction quality issues characterized by blurry outputs(Kingma and Welling, 2013). This blurriness stems from the pixel-wise reconstruction loss, which averages over multiple possible outputs, producing mean-like reconstructions rather than sharp, realistic images(Larsen et al., 2016). Hierarchical architectures offer potential improvements by introducing multiple levels of abstraction in the latent representation, allowing for more nuanced feature learning even at low resolutions.

1.2 Problem Statement

Most research on facial image generation targets high-resolution outputs (128×128 pixels or higher), requiring substantial computational resources and large-scale datasets (Karras et al., 2020). This focus creates a knowledge gap regarding generative model behavior when computational resources are severely limited. Specifically, there is insufficient empirical evidence on how hierarchical VAE architectures perform for low-resolution face generation where facial features must be preserved within a 32×32 pixel constraint. Furthermore, the quality-diversity trade-off inherent to VAEs becomes more pronounced at low resolutions, necessitating careful analysis of perceptual quality metrics.

1.3 Research Objectives

This research aims to evaluate a Hierarchical VAE implementation for low-resolution face generation with the following specific objectives:

- Implement a four-layer hierarchical VAE architecture optimized for 32×32 face generation.
- Evaluate training dynamics and convergence properties over 600 epochs.
- Quantify reconstruction quality using both pixel-wise and distributional metrics.
- Analyze the fundamental limitations of VAE architectures at low resolution.
- Provide performance benchmarks and identify factors contributing to high FID scores.

1.4 Contribution

This work contributes to the generative modeling literature by:

- Providing empirical evidence of HVAE performance on extremely low-resolution faces
- Demonstrating the feasibility of facial feature preservation at 32×32 resolution
- Establishing quantitative benchmarks including FID score analysis
- Analyzing the blurriness problem and quality-diversity trade-off in VAEs
- Offering a computationally efficient implementation using NumPy for accessibility

2 Literature Review

2.1 Variational Autoencoders

Variational Autoencoders, introduced by Kingma and Welling (2013), provide a principled probabilistic framework for generative modeling. The VAE objective combines reconstruction accuracy with latent space regularization through the Evidence Lower Bound (ELBO):

$$L = E[log\ p(x|z)] - KL(q(z|x)\ ||\ p(z))$$

The first term encourages accurate reconstruction, while the KL divergence term regularizes the approximate posterior toward a prior distribution, typically N(0, I). This dual objective provides training stability but often results in blurry reconstructions due to the pixel-wise reconstruction loss averaging over multiple possible outputs (Dosovitskiy and Brox, 2016).The blurriness problem has been extensively documented in VAE literature, with Zhao et al. (2019) identifying it as a fundamental consequence of maximum likelihood training with Gaussian decoders.

2.2 The Blurriness Problem in VAEs

The characteristic blurriness of VAE reconstructions stems from the choice of reconstruction loss. Mean Squared Error (MSE) loss, commonly used in VAEs, penalizes pixel-wise deviations equally, leading the model to generate outputs that represent the average of multiple plausible reconstructions rather than committing to specific sharp details (Mathieu et al., 2015). Larsen et al. (2016)demonstrated that combining VAE objectives with adversarial training can improve perceptual quality, while (Tolstikhin et al., 2017) proposed Wasserstein autoencoders as an alternative that can achieve sharper reconstructions.
Dai and Wipf (2019) analyzed the VAE objective and showed that the reconstruction term's dominance during training, combined with the independence assumption across pixels, fundamentally limits the sharpness of generated images. This effect is amplified at low resolutions where individual pixel values carry more semantic weight.

2.3 Hierarchical Latent Representations

Hierarchical VAEs extend the standard VAE framework by introducing multiple levels of latent variables, allowing for more expressive representations(Sønderby et al., 2016). The hierarchical structure enables the model to learn features at different levels of abstraction. Vahdat and Kautz (2020) demonstrated that hierarchical architectures with multiple stochastic layers achieve superior performance on image generation tasks compared to single-layer VAEs, particularly for complex datasets like CelebA-HQ.
(Maaløe et al., 2019) introduced auxiliary variables in hierarchical VAEs to improve the expressiveness of the approximate posterior, addressing posterior collapse issues. However, they note that even with hierarchical structures, the fundamental trade-off

between reconstruction accuracy and latent space regularization persists, affecting perceptual quality

2.4 Evaluation Metrics for Generative Models

The Fréchet Inception Distance (FID), introduced by (Heusel et al., 2017), has become the standard metric for evaluating generative model quality. FID measures the distance between feature distributions of real and generated images extracted from a pre-trained Inception network. Lower FID scores indicate better perceptual quality and diversity. However, Chong and Forsyth (2020) highlight that FID scores are resolution-dependent and that scores above 100 typically indicate poor perceptual quality or insufficient diversity.
Borji (2022) provides a comprehensive review of generative model evaluation metrics, noting that FID correlates well with human judgment but can be misleading when comparing models across different resolutions or datasets. For low-resolution images, the lack of high-frequency details inherently limits achievable FID scores.

2.5 Low-Resolution Face Generation

Face generation research has predominantly focused on high-resolution outputs. The CelebA dataset (Liu et al., 2015) has become a standard benchmark, typically used at resolutions of 64×64 or higher. Compressed representations at 32×32 resolution present unique challenges in preserving identity-defining features such as eye positioning, facial structure, and basic feature placement.
(Nash et al., 2021) investigated the resolution dependence of generative models and found that VAEs struggle disproportionately at low resolutions compared to GANs, primarily due to the blurriness issue. However, they note that VAEs maintain better mode coverage, generating more diverse outputs even when perceptual quality is limited.

2.6 Research Gap

While hierarchical VAEs have shown promise for complex image generation, their application to extremely low-resolution facial images remains underexplored. The computational efficiency of low-resolution generation makes it relevant for edge deployment, privacy applications, and resource-constrained scenarios, yet empirical benchmarks are lacking. Furthermore, there is limited analysis of how the fundamental VAE limitations manifest at 32×32 resolution and what FID scores are achievable under these constraints.

3 Methodology

3.1 Dataset and Preprocessing

This study utilized the CelebFaces Attributes (CelebA) dataset. To align with the computational constraints of this research, a curated subset of 5,000 images was used. Each image was preprocessed by resizing to a 32×32×3 resolution using a Lanczos filter to maximize feature preservation during downscaling. Pixel values were normalized to the [0, 1] range, converted to float64 for numerical stability, and a robust error-handling routine was implemented to skip any corrupted files.

3.2 Model Architecture

A symmetric, fully-connected Hierarchical VAE was designed with a four-layer encoder-decoder structure to learn compressed facial representations progressively.

Encoder: The encoder maps an input image $x \in \mathbb{R}^{32\times32\times3}$ to a latent distribution through progressive dimensionality reduction: Flatten(3072) → Dense(512) → Dense(256) → Dense(128) →$[\mu(128), \log\sigma^2\ (128)]$. All layers use tanh activation for stable gradients.

Reparameterization: The latent vector is sampled using the reparameterization trick $z=\mu+\sigma\odot\varepsilon$ where $\varepsilon\sim N(0,1)$. This allows gradient-based learning through the stochastic sampling step.

Decoder: The decoder mirrors the encoder, reconstructing the image from the latent code through symmetric expansion: Dense(128) → Dense(256) → Dense(512) → Dense(3072) → Sigmoid → Reshape(32×32×3). A sigmoid activation ensures output pixels are in the valid [0, 1] range.

3.3 Training Configuration and Loss

The model was trained for 600 epochs with a batch size of 16 using gradient descent. Key hyperparameters included a learning rate of 0.003 and He initialization. Adopting a β-VAE formulation , the KL weight (β) was set to 0.01 to prioritize reconstruction quality over strict latent space regularization. The loss function combines:

$$L_{total}=L_{recon}+\beta\cdot L_{KL}$$

where L_{total} is the Mean Squared Error (MSE) and L_{KL} is the Kullback-Leibler divergence between the learned latent distribution and the prior $N(0,1)$

3.4 Evaluation Metrics

Model performance was quantified using three metrics:

- Reconstruction Loss: The pixel-wise MSE, normalized by input dimensionality (3072).
- KL Divergence: Measures how closely the latent distribution matches the Gaussian prior.
- Approximate FID Score:

$$FID \approx ||\mu_r - \mu_g||_2^2 + Tr(\sum_r + \sum_g)$$

A computationally efficient approximation of the Fréchet Inception Distance, calculated from the statistics of pixel distributions to assess the perceptual quality and diversity of generated images compared to the real data.

4 Implementation Details

The model was implemented in Python 3.13 using a minimalist, NumPy (numerical computation),Pandas (metadata handling), with PIL for image processing. This deliberate choice of a NumPy-only implementation, which resulted in a 2.4-hour CPU training time, prioritized four key objectives: accessibility (no GPU or complex frameworks required), transparency (all operations are explicit), educational value (clear demonstration of VAE mechanics), and numerical stability (using float64 precision).

The codebase was architected into modular components—DataPreprocessor, HierarchicalVAE, Trainer, and Generator—to systematically handle data loading, model definition, the training loop, and evaluation, respectively. This modular design ensures the system is both robust and easily modifiable for further experimentation.

5 Results

5.1 Training Dynamics and Final Metrics

The model demonstrated stable convergence over 600 epochs, progressing through three distinct phases: a rapid initial descent, a period of gradual refinement, and a final convergence plateau. The final evaluation metrics were:

- Reconstruction Loss: 0.0726 (indicating reasonable pixel-wise accuracy).
- KL Divergence: 0.5056 (signifying moderate latent space regularization).
- FID Score: 170.23 (revealing a significant perceptual quality gap).

5.2 Analysis of the High FID Score

The FID score of 170.23 is notably high, attributed to several compounding factors:

- Fundamental Limitations: The inherent blurriness of VAEs and the 32×32 resolution constraint.
- Implementation Factors: The use of a simplified, pixel-based FID approximation and training on a limited subset of 5,000 images.

5.3 Qualitative & Loss Analysis

Qualitative assessment confirmed systematic blurriness and a tendency towards "average" faces, though basic facial structure and color were preserved. Analysis of the loss components showed the KL divergence initially decreased as the model learned the latent space, then increased as regularization took effect, ultimately stabilizing. The low β value (0.01) prioritized reconstruction, explaining the controlled KL divergence and contributing to the observed blurriness.

6 Discussion

6.1 Interpretation of Results and The High FID Score

The results demonstrate a clear dichotomy: the low reconstruction loss (0.0726) confirms the hierarchical architecture successfully learned compressed facial representations, while the high FID score (170.23) exposes the fundamental disconnect between pixel-wise accuracy and perceptual quality in VAEs (Dosovitskiy and Brox, 2016). This score is contextualized by several factors: the inherent blurriness of MSE-based VAEs, the challenging 32×32 resolution which mismatches standard FID feature extractors, the use of an approximate pixel-based FID calculation, and the constrained training scale (5,000 images).

6.2 The Blurriness Problem and Architectural Evaluation

The characteristic blurriness is identified as a direct consequence of the pixel-wise MSE loss, which forces the model to output the "average" of possible reconstructions. The discussion notes potential solutions from literature, such as using perceptual loss or adversarial training. Despite this, the hierarchical architecture itself was effective, enabling stable training and gradual feature abstraction without instability.

6.3 Limitations and implications

When compared to literature, the results are consistent with known VAE performance at low resolutions(Larsen et al., 2016). The study acknowledges key limitations, including the simplistic fully-connected architecture, the resolution constraint, and the approximate evaluation metrics. Nonetheless, its implications are valuable: it provides an empirical baseline for low-resolution VAE performance and proves the feasibility of conducting meaningful generative modeling experiments on standard CPU hardware, offering significant educational and benchmark value.

7 Conclusion

7.1 Research Summary and Key Findings

This research investigated Hierarchical Variational Autoencoder performance for low-resolution face generation at 32×32 pixel resolution. The implemented four-layer encoder-decoder architecture with progressive dimensional reduction [512, 256, 128, 128] demonstrated stable training over 600 epochs, achieving final reconstruction loss of 0.0726, KL divergence of 0.5056, and FID score of 170.23. These results provide empirical evidence that hierarchical VAEs can learn meaningful facial representations under severe computational constraints, while highlighting the fundamental limitations of VAE architectures for perceptual quality.
The high FID score of 170.23 and characteristic image blurriness reflect well-documented VAE limitations that become pronounced at low resolution. Analysis reveals that pixel-wise MSE reconstruction loss, resolution constraints, and the quality-diversity trade-off inherent to VAEs collectively contribute to limited perceptual quality despite reasonable pixel-wise reconstruction accuracy

7.2 Successful Objective Fulfillment

The research met all its stated objectives:

- **Implementation & Training:** A functional four-layer HVAE was developed and trained stably for 600 epochs.
- **Comprehensive Evaluation:** Convergence dynamics were analyzed, and performance was quantified using multiple metrics (Reconstruction Loss: 0.0726, KL Divergence: 0.5056, FID: 170.23).
- **Critical Analysis:** The systematic limitations of VAEs at low resolution, particularly blurriness and high FID, were thoroughly investigated.
- **Benchmark Establishment:** The results provide a empirical baseline for future work on resource-constrained generative models.

7.3 A Path Forward for Improvement

To overcome the identified limitations and significantly improve perceptual quality, the following strategies are recommended:

- **Architectural Innovations:** Transition to convolutional layers with residual connections to capture spatial hierarchies. Incorporating multi-scale processing and attention mechanisms can further enhance feature preservation.
- **Enhanced Loss Functions:** Replace or augment the pixel-wise MSE loss with perceptual (VGG-based) and adversarial losses to drive the model towards sharper, more realistic outputs.
- **Refined Training Strategies:** Employ progressive training, β-annealing schedules, larger batch sizes, and advanced optimizers like Adam to improve convergence and model quality.

- **Robust Evaluation:** Future evaluations should use standard feature-based FID, incorporate human perceptual studies, and include additional metrics like Inception Score and Precision/Recall for a complete assessment.

7.4 Future Research Directions

This work opens several promising avenues for future investigation:

- **Hybrid Models:** Exploring VAE-GAN hybrids or Vector-Quantized VAEs (VQ-VAE) to combine the stability of VAEs with the sharpness of GANs.
- **Alternative Frameworks:** Assessing other generative models like diffusion models or normalizing flows for their efficiency-quality trade-off at low resolutions.
- **Theoretical and Applied Work:** Investigating the fundamental limits of low-resolution generation and tailoring models for specific applications like mobile deployment or privacy-preserving data synthesis.

7.5 Concluding Remarks

This research addresses a significant gap in generative modeling literature by demonstrating that hierarchical VAE architectures can learn facial representations at extremely low resolutions, while providing critical analysis of the limitations encountered. The reconstruction loss of 0.0726 demonstrates pixel-wise accuracy, but the FID score of 170.23 and characteristic blurriness reveal the disconnect between pixel-level metrics and perceptual quality—a fundamental challenge in VAE research.

The results contextualize VAE performance at 32×32 resolution, establishing baseline metrics and identifying specific factors contributing to quality limitations. While the achieved perceptual quality remains limited, the stable training dynamics and computational accessibility make this approach valuable for educational contexts and as a foundation for more sophisticated architectures.

The empirical benchmarks and critical analysis provided through this work contribute to understanding the feasibility and limitations of resource-constrained generative modeling, offering insights relevant for edge deployment, privacy applications, and scenarios where computational efficiency must be balanced against perceptual quality.

References

1. BORJI, A. 2022. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding,* 215**,** 103329.
2. CHONG, M. J. & FORSYTH, D. Effectively unbiased fid and inception score and where to find them. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020. 6070–6079.
3. DAI, B. & WIPF, D. 2019. Diagnosing and enhancing VAE models. *arXiv preprint arXiv:1903.05789*.

4. DOSOVITSKIY, A. & BROX, T. 2016. Generating images with perceptual similarity metrics based on deep networks. *Advances in neural information processing systems,* 29.
5. HEUSEL, M., RAMSAUER, H., UNTERTHINER, T., NESSLER, B. & HOCHREITER, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems,* 30.
6. HO, J., JAIN, A. & ABBEEL, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems,* 33**,** 6840–6851.
7. KARRAS, T., AITTALA, M., HELLSTEN, J., LAINE, S., LEHTINEN, J. & AILA, T. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems,* 33**,** 12104–12114.
8. KARRAS, T., LAINE, S. & AILA, T. A style-based generator architecture for generative adversarial networks. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019. 4401–4410.
9. KINGMA, D. P. & WELLING, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
10. LARSEN, A. B. L., SØNDERBY, S. K., LAROCHELLE, H. & WINTHER, O. Autoencoding beyond pixels using a learned similarity metric. International conference on machine learning, 2016. PMLR, 1558–1566.
11. LIU, Z., LUO, P., WANG, X. & TANG, X. Deep learning face attributes in the wild. Proceedings of the IEEE international conference on computer vision, 2015. 3730–3738.
12. MAALØE, L., FRACCARO, M., LIÉVIN, V. & WINTHER, O. 2019. Biva: A very deep hierarchy of latent variables for generative modeling. *Advances in neural information processing systems,* 32.
13. MATHIEU, M., COUPRIE, C. & LECUN, Y. 2015. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*.
14. NASH, C., MENICK, J., DIELEMAN, S. & BATTAGLIA, P. W. 2021. Generating images with sparse representations. *arXiv preprint arXiv:2103.03841*.
15. SØNDERBY, C. K., RAIKO, T., MAALØE, L., SØNDERBY, S. K. & WINTHER, O. 2016. Ladder variational autoencoders. *Advances in neural information processing systems,* 29.
16. TOLSTIKHIN, I., BOUSQUET, O., GELLY, S. & SCHOELKOPF, B. 2017. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*.
17. VAHDAT, A. & KAUTZ, J. 2020. NVAE: A deep hierarchical variational autoencoder. *Advances in neural information processing systems,* 33**,** 19667–19679.
18. ZHAO, S., SONG, J. & ERMON, S. Infovae: Balancing learning and inference in variational autoencoders. Proceedings of the aaai conference on artificial intelligence, 2019. 5885–5892.