Realistic Facial Attribute Control via Variational Autoencoding

Bogdan CĂPRIȚĂ and Stelian SPÎNU

***Abstract*—This paper presents the design and implementation of a generative system based on a variational autoencoder (VAE) for reconstructing and manipulating facial features in images. Positioned within the field of artificial intelligence, the work explores the latent space of a convolutional VAE trained on the CelebA dataset, which includes over 200,000 images labeled with 40 binary facial attributes. The pipeline involves preprocessing, geometric alignment using Dlib and OpenCV, encoder-decoder design, and GPU-accelerated training. A key innovation is the composite loss function, combining Binary Cross-Entropy, perceptual loss from VGG19 activations, and Kullback-Leibler divergence. This improves reconstruction fidelity by preserving local detail, global structure, and latent distribution regularity. To further enhance control, the paper introduces a selective orthogonalization of latent vectors using the Gram-Schmidt process, informed by correlation matrix analysis. This reduces attribute entanglement and enables clearer, more independent manipulation of facial features. The results show high-quality reconstructions and consistent semantic edits, highlighting the model's effectiveness in deepfake-inspired applications. The system provides a foundation for further research in visual content generation, while also addressing controllability and interpretability within generative models.**

***Index Terms*—facial editing, latent space, orthogonalization, variational autoencoder.**

I. Introduction

Recent advancements in generative modeling have enabled significant progress in facial image synthesis, editing, and transformation. Among these, variational autoencoders (VAEs) [1] have emerged as a promising alternative to generative adversarial networks (GANs) [2], particularly for tasks requiring interpretability and controlled manipulation of semantic features. This paper presents the development and evaluation of a VAE-based generative model capable of reconstructing and precisely modifying facial attributes in images. The system encodes facial portraits into a structured latent space, enabling attribute-level transformations while preserving identity and realism.

The proposed solution emphasizes interactive manipulation, allowing users to explore and modify facial features directly. Unlike GANs, which dominate the state-of-the-art but suffer from training instability, high computational requirements, and limited control over latent representations, VAEs offer a more stable and interpretable framework for semantic editing tasks.

This work aims to investigate the potential of VAEs in applications typically addressed by GANs, such as deepfake generation and facial attribute editing. The model is trained on a labeled portrait dataset and optimized using a composite loss function that integrates reconstruction accuracy, latent regularization, and perceptual similarity. The broader objective is to demonstrate that VAEs, when properly designed and optimized, can serve as efficient and controllable generative models suitable for interactive and resource-constrained applications.

The main contributions of this work include the design of a composite loss function that integrates Binary Cross-Entropy, Kullback-Leibler divergence, and perceptual loss from VGG19 features, significantly enhancing image reconstruction quality. Furthermore, the paper introduces a method for extracting representative vectors associated with facial attributes in the latent space. By analyzing the correlation matrix and applying the Gram-Schmidt orthogonalization process, the method ensures reduced entanglement and more interpretable, independent control over attribute modifications.

To encourage transparency and reproducibility, the full implementation of the proposed method is available at https://github.com/BogdanCPR/Facial-Attribute-Control-via-VAE.

II. Related Work

Facial attribute editing has been an active area of research in computer vision and generative modeling. Recent methods have largely focused on leveraging the power of generative adversarial networks to modify specific facial attributes while preserving realism and identity. This section reviews several representative approaches and highlights the differences between these models and the VAE-based method proposed in this paper.

AttGAN [3] introduced an architecture that allows editing only the desired facial attributes while keeping other features unchanged. It achieves this by combining an attribute classification constraint with a reconstruction loss and adversarial training. AttGAN demonstrated that enforcing strict disentanglement is not necessary if attribute supervision is handled effectively. However, the reliance on GAN training makes it sensitive to hyperparameters and computationally demanding.

PA-GAN [4] proposed a progressive attention mechanism that enhances attribute manipulation by focusing on the most relevant facial regions during generation. The model improves attribute transfer fidelity and reduces unintended changes in non-target areas. Nevertheless, its multi-stage adversarial training process introduces complexity and requires fine-tuning to maintain stability.

MU-GAN [5] employed a multi-attention mechanism to

B. CĂPRIȚĂ is with the Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: bogdan.caprita@mta.ro).

S. SPÎNU is with the Military Technical Academy "Ferdinand I", Bucharest, Romania (e-mail: stelian.spinu@mta.ro).

allow the network to adaptively locate and process multiple regions of interest for editing. This approach enhances flexibility for multi-attribute manipulation, but again depends on adversarial training and lacks interpretability in the latent space.

CAFE-GAN [6] introduced a complementary attention mechanism that allows arbitrary facial attribute manipulation by learning disentangled features through attention-based modules. While this approach offers strong performance and generalizability, it is computationally intensive and less suitable for lightweight or interactive applications.

In contrast to these GAN-based methods, the approach proposed in this paper leverages a Variational Autoencoder (VAE) framework, which provides a more interpretable and stable training process. By avoiding adversarial components, our model enables more controlled semantic manipulation of facial features. Moreover, the use of a composite loss function and an orthogonalization strategy for attribute vector extraction facilitates more precise and disentangled editing capabilities, making it better suited for applications requiring transparency and user interactivity.

III. METHODS

A. Dataset

A crucial aspect in the development of the proposed system was the selection of a suitable dataset that supports both the training efficiency and the semantic objectives of a variational autoencoder. Due to the high number of trainable parameters and the complex nature of generative latent modeling, a large and well-annotated dataset is essential to prevent underfitting and to ensure strong generalization performance on unseen data.

To enable precise and controllable facial attribute editing, the dataset must also provide structured labels indicating specific facial features. These labels are critical for identifying target attributes, guiding latent manipulation, and quantitatively evaluating the model's editing capabilities.

Based on these criteria, the CelebA dataset [7] was selected. Created by researchers at the Chinese University of Hong Kong, CelebA is widely adopted in the academic community for its diversity and rich annotations. It contains 202,599 color images of 10,177 unique celebrity identities, each labeled with 40 binary facial attributes, including expressions, structural traits, accessories, and hairstyles.

To standardize the training data, all images were aligned and resized using facial landmarks provided by CelebA (e.g., eye centers, nose tip, mouth corners). This preprocessing step ensures consistent facial positioning, thereby facilitating stable and structured learning of facial components such as eyes, nose, and mouth. The same preprocessing is applied to any external input images to maintain compatibility with the trained model and ensure reliable reconstruction and editing results.

B. Preprocessing of External Input Images

To ensure compatibility with the variational autoencoder trained on the CelebA dataset, external input images must undergo a preprocessing pipeline, as illustrated in Fig. 1. This process guarantees that facial features are properly aligned and standardized prior to encoding. The first step involves converting the original image to grayscale, reducing computational complexity for face detection. A single-face verification step is then applied using the frontal face detector from the Dlib library [8], which relies on a pre-trained convolutional neural network capable of identifying faces under various lighting and pose conditions.

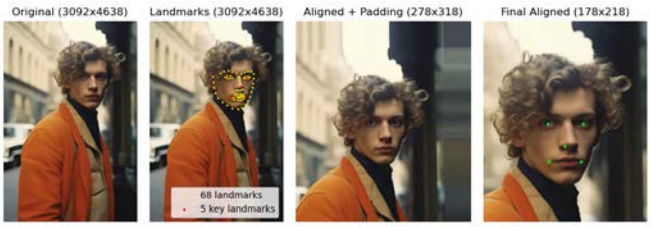

Figure 1. Visualization of the alignment and resizing process

Once a face is detected, Dlib's 68-point facial landmark predictor is used to localize key facial features. From these, five essential points are extracted: the centers of the eyes, the nose tip, and the corners of the mouth. These points are used to compute a 2D affine transformation matrix that aligns the face to a standard orientation. The target reference coordinates were computed as the average positions of the five key landmarks across the CelebA dataset. Using these and the detected landmarks, a geometric transformation comprising translation, rotation, and scaling is applied using OpenCV [9], aligning the input face to the standard CelebA format. To prevent cropping or distortion during this step, a 50-pixel padding is added around the original image.

After alignment, the resulting image is cropped to CelebA's standard size of 178×218 pixels. In cases where affine transformations introduce black borders or incomplete areas, an inpainting algorithm is applied to reconstruct missing regions using surrounding visual information. The final step involves resizing the processed image to 128×128 pixels, which matches the input resolution expected by the VAE.

This preprocessing pipeline ensures that external images are geometrically and semantically consistent with the training data, enabling reliable reconstruction and attribute editing through the autoencoder model.

C. Model Architecture

The proposed model follows the classical structure of a variational autoencoder, consisting of two deep convolutional neural networks: an encoder and a decoder. The encoder compresses input images into a compact probabilistic latent representation, while the decoder reconstructs images from latent vectors.

Encoder

The encoder receives input RGB images of size 128×128 pixels and processes them through a sequence of four convolutional blocks. Each block consists of a 2D convolutional layer with a 4×4 kernel and stride 2, followed by a Leaky ReLU activation, batch normalization, and dropout (rate 0.2). The number of filters increases progressively (64, 128, 256, 512), halving the spatial resolution at each stage while capturing increasingly abstract features. These layers enable the network to learn hierarchical facial representations.

 Digital Object Identifier 10.32754/JMT.2026.1.06

The output is flattened into a 1D tensor, passed through a dense layer of 512 units with batch normalization and dropout, forming a compact global representation. Finally, two separate dense layers generate vectors for the mean (μ) and the log-variance $(\log\sigma^2)$, each of size 512, defining a Gaussian distribution over the latent space.

Unlike deterministic autoencoders, VAEs learn to model a probability distribution over the latent variables. To enable gradient-based optimization, the reparameterization trick is applied. A standard Gaussian noise vector ε is sampled and transformed using the learned parameters, according to the equation:

$$z = \mu + \sigma \odot \varepsilon \quad (1)$$

where $\odot$ denotes element-wise multiplication. This transformation ensures that the sampling operation is differentiable, allowing the model to be trained via backpropagation.

The overall structure of the encoder, including convolutional layers and the latent distribution output, is summarized in Fig. 2.

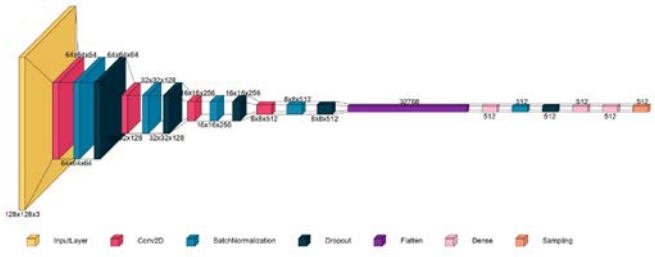

Figure 2. Encoder architecture

Decoder

The decoder reconstructs images from the latent vector $z \in \mathbb{R}^{512}$. It begins with a dense layer that projects the latent vector into a tensor of shape 8×8×512, followed by a Leaky ReLU activation and reshaping into a 3D feature map. Image reconstruction is performed through a sequence of four Conv2DTranspose layers, each doubling the spatial resolution (from 8×8 to 128×128) while progressively reducing the number of filters (256, 128, 64, 32). Each deconvolution layer is followed by batch normalization and a Leaky ReLU activation to maintain stable gradients and support efficient learning.

A final Conv2DTranspose layer with 3 filters and sigmoid activation produces the reconstructed image of size 128×128×3, with pixel values normalized between 0 and 1.

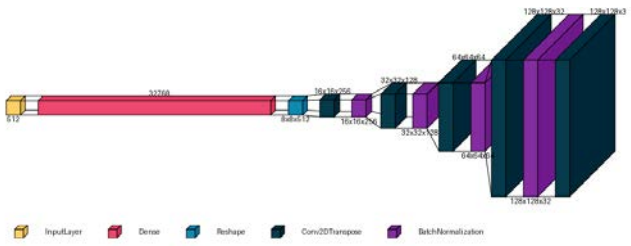

Figure 3. Decoder architecture

The decoder architecture, detailing the upsampling process from latent vector to full-resolution image, is depicted in Fig. 3.

D. Loss Functions

Loss functions play a central role in training variational autoencoders, as they guide the model in learning both a meaningful latent representation and high-quality image reconstructions. Unlike classical VAEs, the implementation proposed in this paper employs a composite loss function composed of three terms: Kullback-Leibler divergence, binary cross-entropy reconstruction loss, and a perceptual loss based on the pre-trained VGG19 network [10].

Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence originates from information theory and quantifies the difference between two probability distributions. In the VAE framework, it measures how closely the latent distribution $q(z|x) = \mathcal{N}(\mu, \sigma^2)$ learned by the encoder aligns with a target standard normal prior $p(z) = \mathcal{N}(0,1)$. The KL divergence is computed using a closed-form expression:

$$D_{KL}\left(\mathcal{N}(\mu, \sigma^2) \,||\, \mathcal{N}(0,1)\right) = \frac{1}{2}\sum_{i=1}^{d}\left(\mu_i^2 + \sigma_i^2 - \log\sigma_i^2 - 1\right) \quad (2)$$

where d is the latent space dimensionality. A scaling factor β is introduced to control the trade-off between reconstruction accuracy and latent space regularization, promoting a balance between realism and semantic structure.

Reconstruction Loss

The Binary Cross-Entropy (BCE) loss is used to measure pixel-wise differences between the input image and its reconstruction. For each image, BCE is computed over all pixels and averaged across the batch. This loss ensures that reconstructions are faithful in terms of pixel values but does not account for high-level perceptual similarity.

To balance the influence of the reconstruction and KL terms, the BCE component is scaled by a factor of 0.1 in the total loss function. This normalization prevents the reconstruction term from overwhelming the regularization effect of the KL divergence.

Perceptual Loss

To further improve the perceptual realism of the reconstructed images, a perceptual loss is introduced. This term compares semantic features extracted from intermediate layers of a pre-trained VGG19 network (trained on ImageNet), which has been widely used for feature-based image comparisons. Four convolutional layers from different blocks of the network are selected, capturing increasingly abstract visual features. Both the input and reconstructed images are passed through the VGG19 feature extractor. The mean squared error (MSE) is computed at each selected layer, and the final perceptual loss is obtained by averaging these values across all layers and images in the batch. Like BCE, this component is also scaled by 0.1 to ensure balance among the three loss terms.

Balancing Reconstruction and Regularization

A key challenge in training VAEs lies in managing the inherent tension between reconstruction fidelity and latent space regularization. While the KL term promotes a smooth,

structured latent space suitable for semantic manipulation and sampling, the reconstruction terms (BCE and perceptual loss) push the encoder to produce input-specific encodings that may deviate from the desired prior. If the KL loss is too dominant, reconstructions degrade; if too weak, the latent space loses structure. Therefore, the training process must strike a careful balance to achieve both visual quality and latent interpretability, which are critical for attribute manipulation tasks.

E. Model Training

The proposed variational autoencoder was trained over 100 epochs using a batch size of 64 and the Adam optimizer with a learning rate of 10^{-4}. These hyperparameters were selected to achieve a balance between training stability and convergence speed. The implementation was carried out using Keras [11] with a TensorFlow backend [12], and training was performed on a GPU-enabled environment. Training progress was monitored on both the training and validation sets. The evolution of each component of the composite loss function: reconstruction loss (BCE), Kullback-Leibler divergence, and perceptual loss, was tracked and recorded independently.

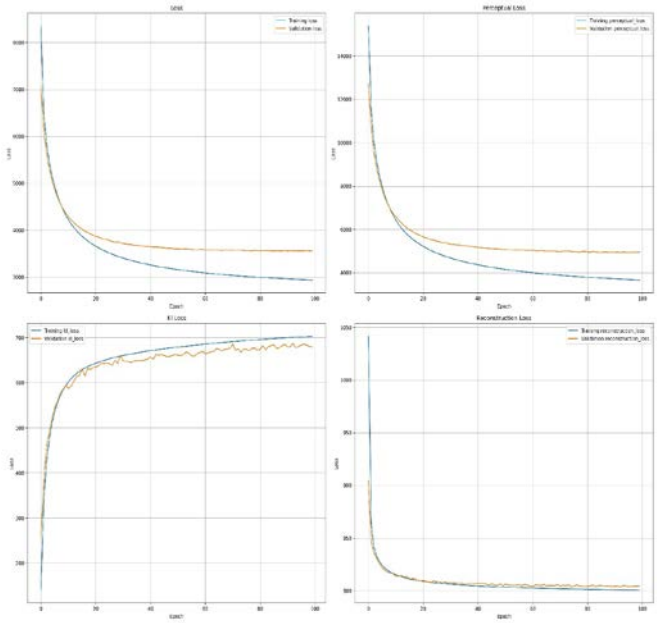

Figure 4. Evolution of loss functions during training and validation

The learning curves (illustrated in Fig. 4) reveal the following dynamics:

• Total loss exhibited a gradual and consistent decrease on the training set, with the validation loss stabilizing around epoch 60–70. This pattern suggests steady convergence and an absence of overfitting.

• The KL divergence displayed a distinctive behavior: it increased gradually during the first phase of training and stabilized later. Initially, the latent distributions $q(z|x) = \mathcal{N}(\mu, \sigma^2)$ remain close to the standard normal prior $p(z) = \mathcal{N}(0,1)$, implying low-information encodings. As reconstruction improves, the encoder is forced to utilize the latent space more effectively, causing the KL divergence to rise. Its eventual stabilization indicates the model has reached a well-balanced trade-off between latent regularization and reconstruction fidelity—an essential condition for effective semantic manipulation in the latent space.

• The reconstruction losses decreased rapidly within the first 20 epochs, after which they entered a plateau phase. This suggests that the model quickly captures the overall facial structure, with later improvements focused on finer visual details.

These dynamics collectively reflect the model's ability to learn meaningful representations while maintaining a stable and well-regularized latent space conducive to controlled facial attribute editing.

F. Facial Attributes in Latent Space

Facial attributes represent a key element in the proposed system, enabling the targeted editing of visual traits in face images. The CelebA dataset provides 40 binary attribute labels per image, corresponding to features such as "Smiling," "Mustache," "Heavy_Makeup," and others. To map these attributes into the latent space learned by the VAE, the entire training set is first encoded into latent representations. For each attribute, the data is split into two groups: one where the attribute is present and another where it is absent. By computing the mean latent vector of each group and taking their difference, a direction vector is derived that reflects the semantic variation corresponding to that attribute. This process is repeated across all 40 attributes, resulting in a set of attribute-specific vectors within the latent space.

To assess the quality and structure of these directions, dimensionality reduction techniques such as t-SNE [13] and UMAP [14] were employed. These methods project the high-dimensional latent directions into a 2D space, enabling visual inspection of the relationships between attributes. The resulting 2D projections, generated via t-SNE and UMAP, are shown in Fig. 5.

The t-SNE visualization reveals coherent clusters among semantically similar attributes—facial hair features like "Mustache," "Goatee," and "Sideburns" are grouped together, as are aesthetic features typically associated with femininity, such as "Wearing_Lipstick" and "Heavy_Makeup."

UMAP projections further reinforce these patterns, while also offering improved separation of clusters and better preservation of global distances. Both projections highlight a strong gender-based separation: masculine features tend to appear in one region, while feminine ones cluster in another. Neutral or ambiguous traits, such as "Smiling" or "Attractive," are often located more centrally, reflecting their relevance across gender categories.

To further explore the relationships between attributes, a Pearson correlation matrix was computed over the set of latent direction vectors. Strong positive and negative correlations reflect co-occurrence or exclusivity between features. For instance, "Male" is strongly negatively correlated with "Wearing_Lipstick" and "Heavy_Makeup," while features like "Young" and "Attractive" show positive association. At the same time, attributes like "Eyeglasses" or "Oval_Face" exhibit near-zero correlation with most others, suggesting semantic independence. While these correlations validate the semantic coherence of the learned latent space, they also reveal significant overlap between some attributes, which can lead to entanglement during manipulation.

 Digital Object Identifier 10.32754/JMT.2026.1.06

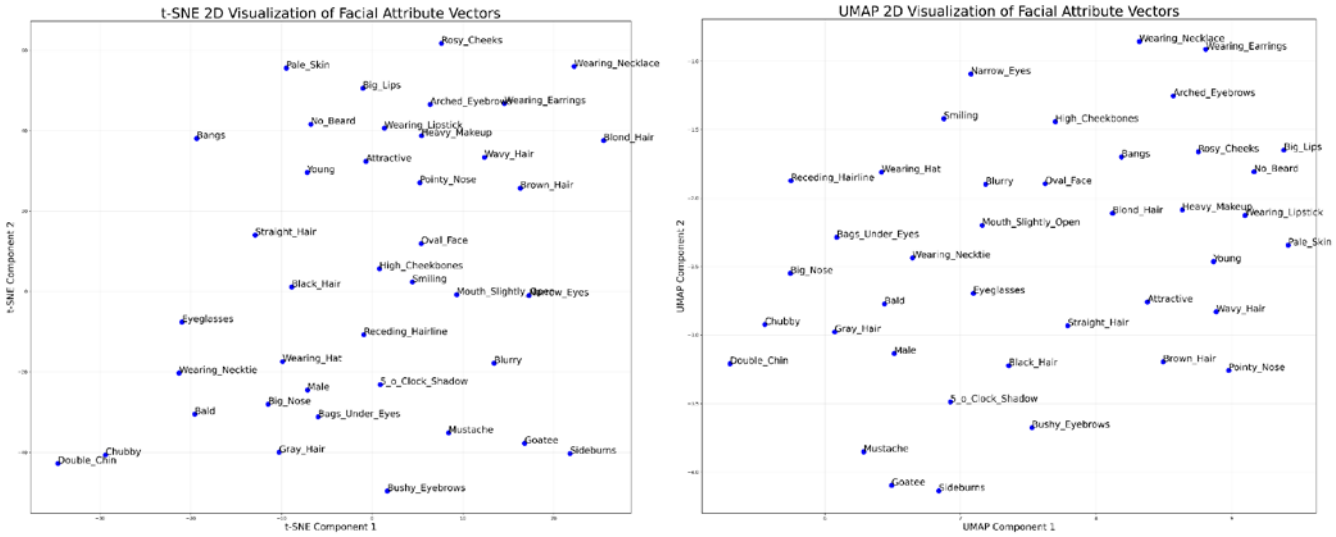

Figure 5. Two-dimensional visualization using t-SNE and UMAP

To address this, a selective orthogonalization strategy was implemented to decorrelate overlapping attribute directions. Based on the correlation matrix, attribute pairs with correlation magnitude in the range $0.6 < |corr(i, j)| < 0.9$ were selected. For each such pair, the vector corresponding to attribute i was orthogonalized with respect to vector j using the Gram-Schmidt method. This involved subtracting the projection of one vector onto the other, thus eliminating their shared component while preserving most of the semantic identity. This operation was applied iteratively, keeping reference vectors fixed to avoid feedback loops. The result was a refined set of attribute directions with lower interdependence, which was confirmed through a second correlation analysis (illustrated in Fig. 6) showing a marked reduction in high-magnitude correlations.

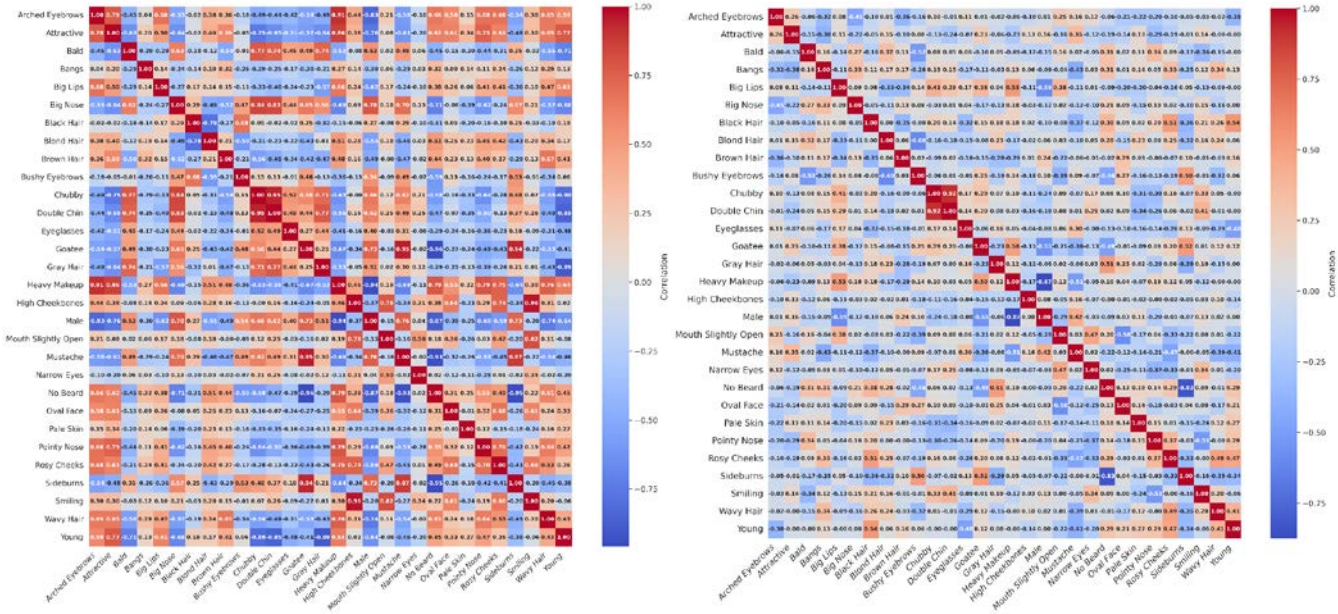

Figure 6. Correlation matrix before and after orthogonalization

With attribute directions defined and refined, facial images can be manipulated in the latent space through vector arithmetic. Given the latent representation z of an input image, one or more attributes can be modified by translating z in the direction of the corresponding vectors. The transformation is described by the equation:

$$z' = z + \sum_{i=1}^{k} \alpha_i v_i \tag{3}$$

where v_i represents the direction vector of the i-th attribute, α_i is a scalar controlling the strength of the modification. This formulation supports the simultaneous adjustment of multiple traits in a single operation. For example, smiling can be added while facial hair is removed and makeup intensified, all by adjusting the latent vector accordingly. The modified latent code z' is then passed through the decoder to reconstruct a new image that reflects the edited attributes.

This entire process, ranging from vector extraction and analysis to orthogonalization and real-time manipulation, demonstrates the effectiveness and flexibility of the

proposed VAE-based system in enabling controllable, interpretable, and high-quality facial attribute editing.

IV. RESULTS

A. Image Reconstruction

Fig. 7 presents a set of reconstructed images from the test dataset, which was completely excluded from both training and validation phases. This separation ensures an unbiased assessment of the model's generalization capabilities and its ability to faithfully reconstruct unseen facial images.

Figure 7. Reconstructions of test images

The results demonstrate that the variational autoencoder successfully preserves the key identity-defining features of each individual. These include face shape, hair color, skin tone, and head orientation, all of which are critical for enabling controlled and realistic attribute editing. Such consistency is essential, as any subsequent manipulation relies on the fidelity of the reconstructed base image.

In terms of background reconstruction, the model shows varied performance. Simple and uniform backgrounds are generally well-reproduced, while more complex or detailed backgrounds tend to be simplified or distorted. This behavior is acceptable, as background fidelity is not a primary objective of the model and does not impact attribute manipulation.

Some limitations are observable in the fine-grained details of the reconstructions. For example, subtle skin textures, eye sharpness, and hair detail may be smoothed or lost. This is a known characteristic of VAE-based architectures, which often prioritize stability and global structure over fine texture, leading to slightly softened outputs.

Despite these limitations, the reconstructions provide a sufficiently accurate foundation for downstream manipulation. As the attribute editing process operates on the reconstructed image, the quality of these outputs directly impacts the realism and precision of semantic modifications. The model's ability to consistently reproduce high-level features makes it well-suited for the intended application.

B. Latent Distribution and Latent Space Regularization

As discussed previously, training a variational autoencoder involves a fundamental trade-off between two competing objectives: ensuring high-fidelity image reconstruction and enforcing a structured latent distribution. The reconstruction loss—composed of binary cross-entropy and perceptual loss—encourages precise encoding of each individual input, which can lead to a disorganized, overfitted latent space. In contrast, the Kullback-Leibler divergence term regularizes the latent space toward a standard Gaussian distribution, favoring semantic structure and facilitating smooth attribute manipulation.

To illustrate the practical implications of this trade-off, three variants of the VAE model were compared, each trained with different weights for the KL term in the composite loss function:

• VAE (Balanced): A model with a moderate KL weight, designed to strike a balance between reconstruction quality and latent regularization.

• VAE Recon+: A model where reconstruction is heavily prioritized, with minimal KL influence.

• VAE KL+: A model where KL regularization dominates, encouraging alignment with the prior distribution at the expense of reconstruction accuracy.

The results are shown in Fig. 8, which compares the reconstruction performance of each variant. The VAE Recon+ model yields the most visually accurate reconstructions, preserving fine details and identity. In contrast, the VAE KL+ model fails to reproduce recognizable faces, due to an overly constrained latent space. The balanced model offers a compromise between both goals, achieving reasonable reconstructions while preserving latent structure.

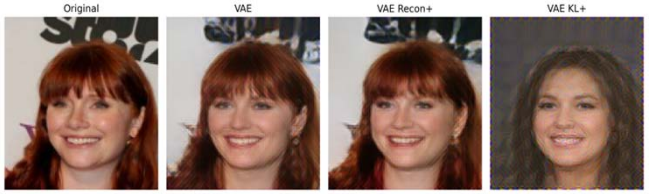

Figure 8. Comparison of reconstructions for three VAE models with different loss term weights

This trade-off becomes even more evident during semantic attribute manipulation, as illustrated in Fig. 9, which shows the effect of modifying the "Smiling" attribute using each model. While VAE Recon+ provides a high-fidelity reconstruction of the original image, it fails to apply the attribute transformation effectively, highlighting a disorganized latent space with poor semantic alignment. In contrast, VAE KL+ exhibits smoother interpolation in the latent space but starts from a degraded image. The balanced model again offers the best overall behavior, supporting both accurate reconstruction and meaningful attribute manipulation.

Figure 9. Modification of the "smiling" attribute for the three model versions

The loss functions used for training each variant were defined as follows:

• VAE (balanced):

Loss = Recon Loss + 10 * KL

• VAE Recon+:

Loss = Recon Loss + 0.1 * KL

• VAE KL+:

Loss = Recon Loss + 100 * KL

These results confirm that an appropriate balance between reconstruction and regularization is essential.

 Digital Object Identifier 10.32754/JMT.2026.1.06

Overemphasizing either component degrades the model's utility, either by compromising reconstruction quality or by eliminating the structure needed for controlled latent manipulation. The balanced configuration ultimately offers the best compromise for the intended application.

C. Manipulating Facial Attributes in Latent Space

A key feature of the proposed VAE-based system is its ability to perform controlled semantic manipulation of facial attributes through vector operations in the latent space. This functionality allows both single-attribute editing and multi-attribute interpolation, enabling fine-grained and realistic transformations of facial images while preserving identity.

Single-Attribute Manipulation

Fig. 10 illustrates an example of modifying the facial expression "Smiling" by adding the corresponding latent direction vector scaled with values ranging from −4 to +4. This technique enables direct control over the intensity of the attribute. Negative values reduce the smile, leading to a neutral expression, while positive values amplify it, producing a wide and expressive smile. Notably, the transformation is physiologically consistent: it affects not only the mouth but also surrounding facial regions such as the cheeks, which appear lifted and more prominent during a strong smile, or relaxed and recessed when the smile is suppressed. This highlights the model's ability to capture realistic co-activations of facial features during expression changes.

Figure 10. Modification of the "smiling" attribute with varying intensities

Multi-Attribute Interpolation

The continuous and semantically organized nature of the latent space also supports the simultaneous modification of multiple facial attributes through linear combinations of their direction vectors. Fig. 11 presents a 2D interpolation grid that combines four attributes: negative "Smiling" (bottom-left), "Bangs" (top-left), "Eyes Closed" (top-right), and "Blond Hair" (bottom-right). Each image in the grid represents a linear blend of the latent directions associated with its respective corners.

Images closer to the center show reduced intensities of all attributes, while those near the corners emphasize one dominant feature. The smooth transitions observed across the grid demonstrate the model's ability to interpolate semantically meaningful transformations while maintaining facial structure and identity. No visual artifacts or inconsistencies are observed, even when multiple attributes are blended simultaneously, reinforcing the coherence and stability of the latent space.

Effect of Attribute Orthogonalization

To improve the independence and specificity of attribute manipulation, a selective orthogonalization process was applied to the latent direction vectors. This step aimed to eliminate overlapping semantic components between highly correlated attributes, potentially leading to cleaner and more isolated modifications.

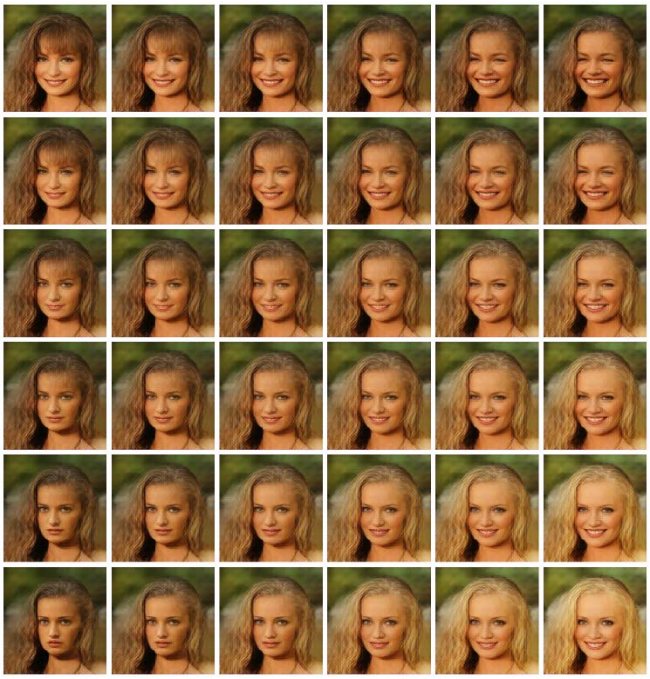

Figure 11. Two-dimensional latent interpolation between four facial attributes

Fig. 12 compares the application of two attributes Bushy Eyebrows and Arched Eyebrows using the original latent direction vectors and their orthogonalized counterparts. In most cases, the visual difference between the two versions is subtle, though occasionally a slightly clearer separation of effects can be observed after orthogonalization. This suggests that the impact of orthogonalization is context-dependent, varying with the specific attribute, image, and latent structure.

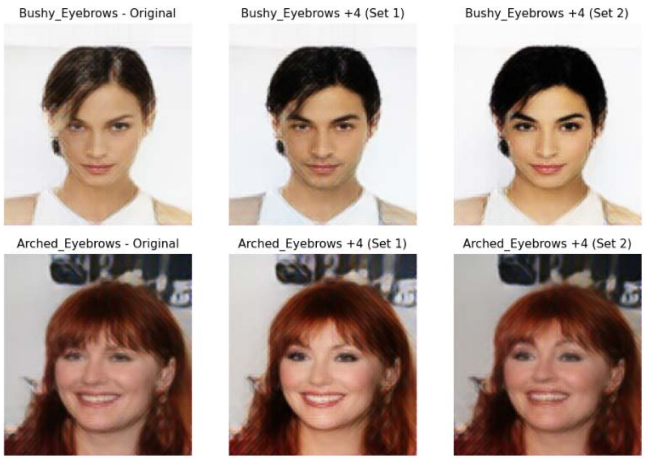

Figure 12. Comparison of results using two sets of attributes: original and orthogonalized

As a result, both versions of the attribute vectors were preserved in the final implementation of the interactive system, giving users the option to select either the original or orthogonalized variant based on the image and desired editing behavior. This design choice reflects the model's flexibility and responsiveness to subtle differences in attribute expression and control.

D. Perceptual Evaluation of Results

To enhance the visual quality of reconstructions, the proposed VAE model incorporates a perceptual loss term based on intermediate feature activations from the pre-

trained VGG19 network. To evaluate the impact of this component, two models were compared: one trained with the perceptual loss, and another using only binary cross-entropy and KL divergence.

Fig. 13 illustrates the difference in reconstruction quality. The model trained without perceptual loss produces blurry and poorly structured faces, with significant identity loss. In contrast, the version incorporating perceptual supervision yields sharper, more coherent reconstructions, preserving both facial structure and identity. These results highlight the effectiveness of perceptual loss in guiding the model toward more semantically faithful image generation.

Figure 13. Comparison of image reconstruction using a VAE with and without perceptual loss

E. Limitations

While the proposed VAE model demonstrates strong flexibility in both reconstruction and facial attribute manipulation, one important limitation must be considered in practical use. This issue is illustrated in Fig. 14, where successive reconstructions are generated by re-encoding and decoding the output of the previous step. Over multiple iterations, a gradual degradation in image quality is observed—details fade, colors become muted, and facial features begin to distort.

Figure 14. Successive reconstructions of the same image

This accumulation of reconstruction error is a known limitation of VAEs, which are not inherently designed for recursive generation. To mitigate this effect, all attribute modifications are applied directly to the original image's latent representation, avoiding chained reconstructions and helping to preserve visual fidelity throughout the editing process.

V. Conclusion

This work explored the use of a variational autoencoder for facial image reconstruction and attribute manipulation, aiming to strike a balance between reconstruction fidelity and latent space structure. The integration of a perceptual loss based on VGG19, alongside the standard reconstruction and KL terms, led to improved visual quality and more coherent identity preservation. Additionally, a selective orthogonalization strategy was introduced to reduce latent vector correlations and enhance the precision of semantic edits. The results confirmed that the model successfully learns a meaningful and structured latent space, enabling smooth and interpretable modifications of facial features.

Although the system performed well, it showed limitations in extreme regions of the latent space and in recursive reconstruction scenarios. Future improvements may involve combining the model with adversarial training to enhance realism, enabling attribute transfer between different faces, and incorporating automatic estimation of existing attribute intensities to support more intuitive and adaptive editing.

References

[1] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *arXiv [stat.ML]*, 2013. doi:10.48550/arXiv.1312.6114

[2] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2672–2680.

[3] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019. doi:10.1109/TIP.2019.2916751

[4] Z. He, M. Kan, J. Zhang, and S. Shan, "PA-GAN: Progressive attention generative adversarial network for facial attribute editing," *arXiv [cs.CV]*, 2020. doi:10.48550/arXiv.2007.05892

[5] K. Zhang, Y. Su, X. Guo, L. Qi, and Z. Zhao, "MU-GAN: Facial attribute editing based on multi-attention mechanism," *arXiv [cs.CV]*, 2020. doi:10.48550/arXiv.2009.04177

[6] J.-G. Kwak, D. K. Han, and H. Ko, "CAFE-GAN: Arbitrary face attribute editing with complementary attention feature," in *Computer Vision – ECCV 2020*, Cham: Springer, 2020, pp. 524–540. doi:10.1007/978-3-030-58568-6_31

[7] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 3730–3738. doi:10.1109/ICCV.2015.425

[8] D. E. King, "Dlib-ml: A machine learning toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[9] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, vol. 25, no. 11, pp. 120–126, Nov. 2000.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR)*, 2015. doi:10.48550/arXiv.1409.1556

[11] F. Chollet *et al.*, *Keras*, https://keras.io, 2015.

[12] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI)*, 2016, pp. 265–283.

[13] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[14] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv [stat.ML]*, 2018. doi:10.48550/arXiv.1802.03426