Journal of Information Hiding and Multimedia Signal Processing ©2026 ISSN 2073-4212
Ubiquitous International Volume **17**, Number **1**, March **2026**

Disassembling Landmark and Texture Decoding for Inverting VGG-Face Embeddings

Ahmad M. Nagm[1]

[1]Department of Computer Engineering and Electronics
Cairo Higher Institute for Engineering, Computer Science and Management
Cairo 11477, Egypt

Ahmed Abdelhafeez[2,3]

[2]Faculty of Computer and Information Technology
Innovation University
Cairo, Egypt

[3]Applied Science Research Centre
Applied Science Private University
Amman, Jordan
ahmed.abdehafeez@iu.edu.eg

Moshira A. Ebrahim[4,*]

[4]Computer Engineering and Information Technology Department
Modern Academy for Engineering and Technology
Cairo, Egypt
mushira.ibrahim@eng.modern-academy.edu.eg

Ayman Al-Ahwal[5]

[5]Faculty of Computers and Information Systems
The Egyptian Chinese University
Egypt
Dr.Ayman.hiet@gmail.com

*Corresponding author: Moshira A. Ebrahim

Received September 16, 2025, revised December 15, 2025, accepted December 17, 2025.

Abstract. *Reconstructing human faces from high-level feature embeddings is a challenging yet essential task in explainable biometrics, face synthesis, and security systems. This paper presents a novel face reconstruction framework that leverages a multi-decoder architecture to invert 4096-dimensional embeddings extracted from the VGG-Face network. Unlike previous approaches, our method decomposes the reconstruction process into two specialized components: a texture decoder and a landmark decoder, each trained independently to improve representation fidelity. By warping the generated texture according to predicted facial landmarks through a differentiable spline transformation, we produce realistic, identity-preserving facial reconstructions. Our pipeline is trained on a curated dataset augmented through morphing and semantic blending to ensure demographic diversity and pose neutrality. Quantitative evaluations using SSIM, PSNR, and Fréchet Inception Distance (FID) demonstrate competitive reconstruction accuracy, while qualitative results confirm high visual plausibility. This modular decoding approach opens new pathways for interpretable and controllable face synthesis from embedding spaces.*

Keywords: Face Reconstruction, VGG-Face, Deep Learning, Texture Decoding, Landmark Detection, Image Warping.

1. **Introduction.** Recent advances in pattern recognition, face recognition, and face-based deep learning techniques predict the face in vector representation form from the complex speech spectrogram [1]. Machine learning can develop automatic face recognition systems to address these applications. On the one hand, recognizing a face is a natural process because people usually do it effortlessly and without much consciousness.

Modern face recognition networks can transform a facial image representing high-level features into an embedded representation or a feature vector that compresses too many values in the image into a compact representation with a reduced number of parameters. These networks opened the scope for great progress in computer vision, which resulted in a great need for perception and interpretation of the output embedded feature vector in order to use it in applications that relate the face to other biometric features using the encoder-decoder architecture. This study focuses on the decoder part, which is based on the work done by [2]. The results imply that many biometric features are correlated with facial features and that the face can be decoded based on them. Many search areas are included in the topic of reconstructing a face from an output-embedded feature vector, some of which we will discuss below.

Deep learning, especially convolutional neural networks (CNNs), has significantly enhanced multimedia identification, and analysis in many computer vision applications in different areas [3, 4, 5]. CNN-based architectures outperform in learning hierarchical and discriminant representations from raw pixel data, outperforming conventional constructed feature techniques. Deep face detection and identification models like VGG-Face, FaceNet, and ArcFace can reliably represent identities in compact embedding spaces despite differences in position, lighting, expression, and occlusion.

The suggested system receives as input face recognition networks such as the Visual Geometry Group (VGG-Face), which converts (224, 224, and 3)-dimensional facial images to a 4096-dimensional facial feature vector. The system then outputs a picture of the human face. When creating facial vectors identical to the input photos and invariant to posture, lightning, and partial occlusion, VGG-face and other neural networks. The following is a summary of this work's primary contributions:

1) Instead of depending just on facial geometry (landmarks) and appearance (texture) at the decoder level, we suggest a fully decoupled face reconstruction system.

on a single generator. Interpretability, stability, and reconstruction fidelity are all enhanced by this explicit decoupling.

2) The suggested design independently trains a landmark decoder and a texture decoder directly from the same 4096-D VGG-Face embedding, in contrast to previous landmark-guided or landmark-conditioned generative models. This allows each module to specialize without shared loss interference.

3) To provide spatial consistency without adversarial training, we present a differentiable spline-based warping stage that geometrically aligns the anticipated texture with the predicted landmarks.

4) We provide a thorough quantitative and ablation analysis that shows the suggested modular architecture works better in SSIM, PSNR, FID, and landmark localization accuracy than single-decoder and non-warping baselines.

This paper is organized as follows: Section 2 describes related work. Section 3 describes the proposed method. Section 4 describes face decoder model. Section 5 presents the training dataset. Section 6 presents the conclusions and future work

2. **Related Work.** Face reconstruction using high-level representations has gained popularity due to its applications in explainable biometrics, privacy analysis, and controlled face synthesis. Modern face recognition algorithms, such as VGG-Face [6], map facial images into compact embedding spaces that maintain identification while removing irrelevant changes like stance, lighting, and occlusion. While these embeddings are extremely discriminative, converting them back into realistic face images is a difficult and ill-posed challenge.

Face reconstruction has several approaches, varying from model-based deep convolutional face autoencoders with a differentiable parametric decoder that analytically encapsulates image construction, where the CNN-based encoder learns to extract semantically meaningful parameters from a single input image for generating facial texture and shape from images by using generative adversarial networks (GANs) to train a vigorous generator of facial texture. Voice-to-face is the branch that generates faces from human voice clips. Several techniques used GANs to enhance realism. For example, a novel facial expression GAN (FE-GAN) [7] is proposed to reconstruct sharper and more discriminative faces taking in consideration emotion and expressions. Two auxiliary emotion classifiers were used to learn more about emotion and identity representations. The results showed that FE-GAN can outperform the previous models in terms of Fréchet inception distance (FID) and inception score (IS) values and generate more realistic face images than previous models. While, MaskGAN [8] was proposed, which has two main components: Dense Mapping Network (DMN), which maps between a free-form and a target image, and Editing Behavior Simulated Training (EBST), which makes the overall framework more robust to various manipulated inputs that dual-editing consistency as the auxiliary supervision signal. The results showed that MaskGAN demonstrated superior performance over other state-of-the-art methods and overcame the drawbacks of operating on a predefined set of face attributes or leaving users little freedom to interactively manipulate images in previous methods.

Landmark detection is an ongoing topic of research that is well established through different approaches that can be roughly classified into three main groups: holistic methods, constrained local model (CLM) methods, and regression-based methods. They vary in the manner in which they use facial appearance and structure information [9]. Many studies have been conducted to address structural consistency using landmark-guided and landmark-conditioned techniques. Modern deep learning approaches [10] use landmarks as supplementary supervision or conditional inputs. However, in the majority of existing frameworks, landmark information is combined with texture creation inside a single network or a jointly optimized loss, resulting in unstable training and confused geometry and appearance.

Other studies investigated the leaking and invertibility of facial descriptors. ID2image [11] demonstrated that face recognition embeddings can partially recover non-identity attributes and landmark positions, allowing for optimization-based inversion with generative models to produce photorealistic reconstructions. Recent approaches, such the Deep Face Decoder [12], recreate deep face templates to visualize embedding spaces. This highlights sophisticated inversion capabilities and issues in retaining geometric purity.

Natural mouth movements with convincing lip sync can create believable video from audio with convincing lip sync, as in [13]. To match the mouth shape with a 3D pose, match the mouth texture based on MFCC audio features for which extracts speaker-specific parameters from the speech, and RNN to achieve even better quality by going directly from raw audio waveforms to mouth shapes or textures. The network can learn to predict emotional states from audio to produce corresponding visuals. In [14] a conditional sequential generative adversarial (CSG) network is suggested with two models: CSG-Emo-Adapted and CSG-Emo-Aware. These models reconstruct orofacial movements from acoustic features, especially happiness. It creates lip motion trajectories that are realistic and flexible. Visual speech recognition (VSR) recognizes speech without audio by using lip motions. Chan et al. [3] used a VGG-M CNN to categorize spoken digits from lip image sequences by extracting discriminative spatial features. Their technique demonstrates the usefulness of CNNs for lip-reading while also resolving issues such as varying speaking speeds and speaker differences.

In addition, a deep learning technique for constructing a real-time 3D facial animation [15] by audio input with low latency is presented. It models the speaking style of a single actor with different genders, accents, or languages, and yields realistic results with low-cost localization, virtual reality avatars, and telepresence. Additionally, an effective deep learning approach [16] is proposed to generate natural-looking speech animation in real time from user speech input. It synchronizes with input speech. It provided machine learning design decisions includeing various animation clips on a variety of characters and voices. It uses a sliding window predictor that simulates input sequences to mouth movements that capture the natural motion. Furthermore, a self-supervised framework (X2Face) [17], was developed for driving face generation using another input face or audio without further network training. This work presented a more robust model than other methods with fewer input data assumptions.

In addition to the autoencoders and GANs, deep CNNs have been used for texture generation combined with multilayer perception for landmarks [2]. Transformer-based decoupled representation learning [18] has been investigated in 3D face reconstruction and dense alignment, with a focus on geometric factor separation for structure estimation. However, these methods frequently require parametric models or complicated priors, which raises computing costs and deviates from straightforward inversion of fixed recognition embeddings.

Image warping is a transformation for mapping positions between one image to another. Image warping has multiple approaches that can be categorized into parametric and non- parametric approaches [19]. In our model, image warping is used in the last stage to apply the landmarks on the predicted face.

Although many of the previous studies use facial information for guiding face synthesis, their architectural function is very different from the suggested method. In landmark-guided GANs, geometry and texture are implicitly entangled inside common layers and landmarks are used as conditional inputs to a single generator. Our system, on the other hand, uses two independently trained decoders that clearly distinguish between appearance synthesis and geometry prediction. Normalized face synthesis techniques use similar decoupling assumptions, although they frequently rely on common latent spaces or collaborative optimization. In contrast, our approach produces better stability, interpretability, and geometric precision by ensuring architectural independence and re-aligning the outputs only through a differentiable warping stage.

3. The Proposed System.

The system originally consisted of an encoder-decoder architecture. The output embedded feature vector resulting from applying a loss function to the output from a biometric feature encoder such as voice and the feature vector from the pre-last layer of the VGG-Face of the same dimensions is then entered as input to the face decoder. To extract the features from the input image, we utilized the VGG-Face model, a face recognition model pre-trained on a large-scale face dataset [20] and extracted a 4096-D face feature vector from the pre-last layer without activation. Profile-based face recognition encounters many challenges: inconsistent postures, differing illumination, blockage, and facial expressions [21]. The same challenges are present for face reconstruction, but these face features have been proven to carry sufficient data to reconstruct the corresponding faces while being vigorous to these variations [1]. The face recognition module takes frames that have human faces in them, extracts their physical facial features, and returns these features in a feature vector. The face recognition module is a pre-trained and fixed component that we used in our model to extract the face features and store them in vectors. Its output is considered as the true label that is compared with the audio encoder to minimize the cost function. Eventually, after the training is completed and the loss decreases to an acceptable level, this component is thrown away. The VGG Face Network is a CNN used to produce a 4096-feature vector for all images or tensors of the training dataset. The original architecture of the system is similar to the work done by [1], but the decoder has a novel architecture for both the landmarks and texture decoders.

4. Face decoder model.

4.1. System modules.

The idea of separating the texture and landmarks for modeling the face has proved to be more robust to variations. The 'texture' is essentially the lineaments of the face without the impact of the facial features' shapes and their relative positions. The texture is generated by deforming all the faces in the training set to have a structure similar to that of the mean face. Texture and landmarks are used in conjunction to represent the general appearance of each face [2]. The input image is separated during training, the into landmarks by a landmark detection algorithm and a texture by applying the mean landmarks. It is then entered as an input to the VGG face, and a 4096-D vector is extracted from the pre-last layer without activation. This feature vector is the input for both the multilayer perception network (MLP) and the convolutional network.

The output of the MLP is the predicted landmarks, which is a 68x2 matrix containing the locations of 68 specific points in the face. The output of the convolutional neural network (DCNN) is 224x224x3 which is the predicted texture of the input image. The loss function for the landmarks model is the mean squared error, where the landmarks extracted from the image are considered the ground truth, while the loss function for the texture model is the mean absolute error, where the texture extracted from the original image is considered the label.

The predicted landmarks are applied to the predicted texture through differentiable warping resulting in the output image. During testing, the input to the decoder is the output from the biometric encoder, which is a 4096-D feature vector that is then entered into the MLP and DCNN as input. A dashed rectangle marks the testing as shown in Figure 1.

4.2. Landmarks model.

The landmark model predicts the location of the 68 key points in the face. As shown in Figure 2, these points represent the shape, emotions, and orientation of the face. Making a dedicated model for learning landmark positions allows us to separately penalize the error in detecting the high- level features that represent the state of the face by predicting only 68 points.

The reason behind using a landmark model is to ensure that the predicted face has proper landmark positions. These positions control the deformation of the predicted faces. Therefore, it is crucial to have minimum loss in the positions of the landmark points to make the predicted faces as visually appealing and realistic as possible. In addition, creating a dedicated model for learning landmark positions with no concern about face texture improves the model results in minimizing the loss function.

Facial landmarks are used for localizing and representing facial parts of the human face, such as the eyes, nose, mouth, jaw, chin, eyebrows, and any important key points regarding the human face as shown in Figure 3. The human face can be represented using both the texture and the landmarks of the face. The reason that the data given to the model to train must be of neutral nature and frontal face so that the extracted landmarks differ only on face dimension not due to face movement or face rotation as must

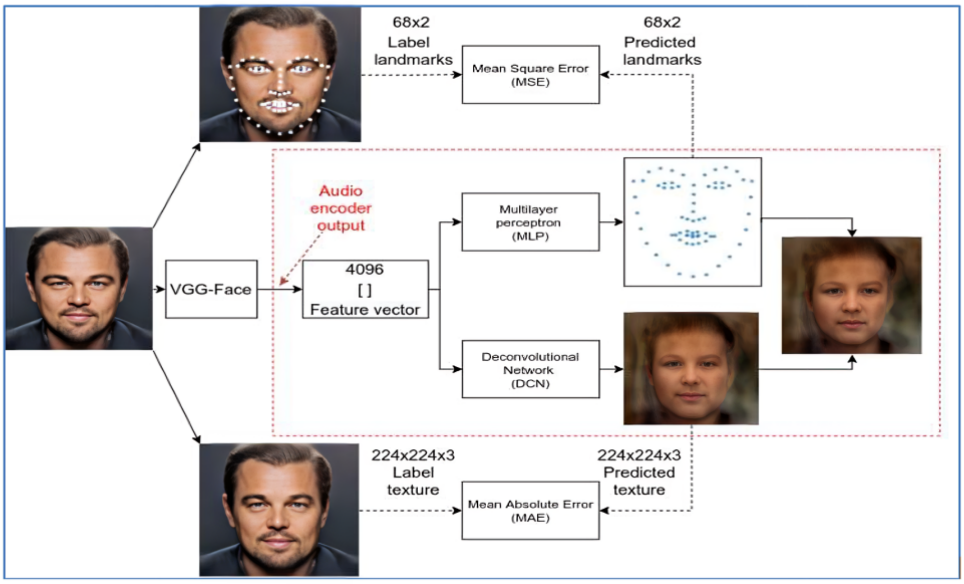

Figure 1. Face decoder architecture.

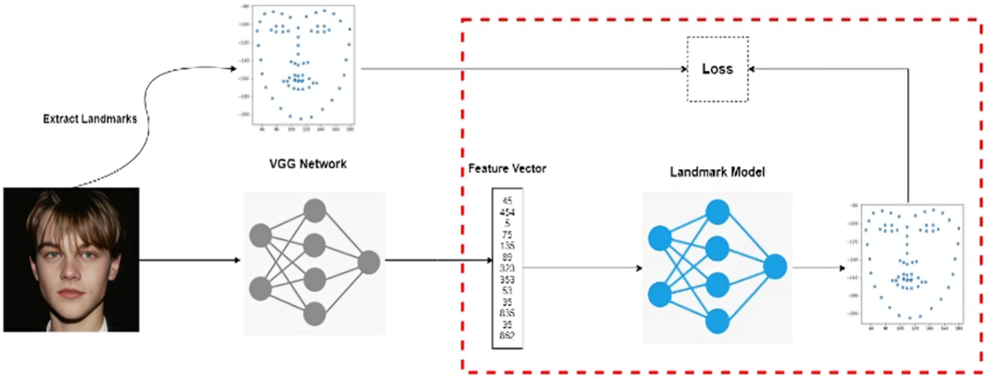

Figure 2. Overview of the landmark model

the only player affecting the landmarks only is the actual shape of the face then the model will be able to generalize well given enough data.

4.3. **The training process.** The optimizer used was the Adam adaptive moment estimation optimizer [22], with an initial learning rate of 0.001, a beta of 0.5, and a decay of 0.000095, as discussed in detail below in the loss section.

The landmark model has 139,994,080 parameters, 5,729,936 trainable parameters, and 134,264,144 non-trainable parameters.

4.4. **Deconvolution Neural Network.** The concept of mirroring the convolutional network to obtain a convolutional network is present in many fields in deep learning, most notably in semantic image segmentation and autoencoders, but it is always integrated with the CNN that follows it in the architecture and is not presented as a separate entity [23]. However, in some applications, a convolutional network of a significantly different reverse architecture than CNN can be used, which gives the same functionality but with a different approach, such as stacking a set of transposed convolutions as shown in Figure 4.

Moreover, for our specific convolutional network architecture, and while sticking to the inverted architecture of the VGG, each pooling layer followed by the convolutional layer is replaced with a transposed convolutional layer, which is in essence the same as the down-sampling spooling layer merged with a convolutional layer.

This architecture is unique in many aspects, including that the DCNN is a stand-alone component that does not complete the CNN as in autoencoders, does not consist of just transposed convolutions

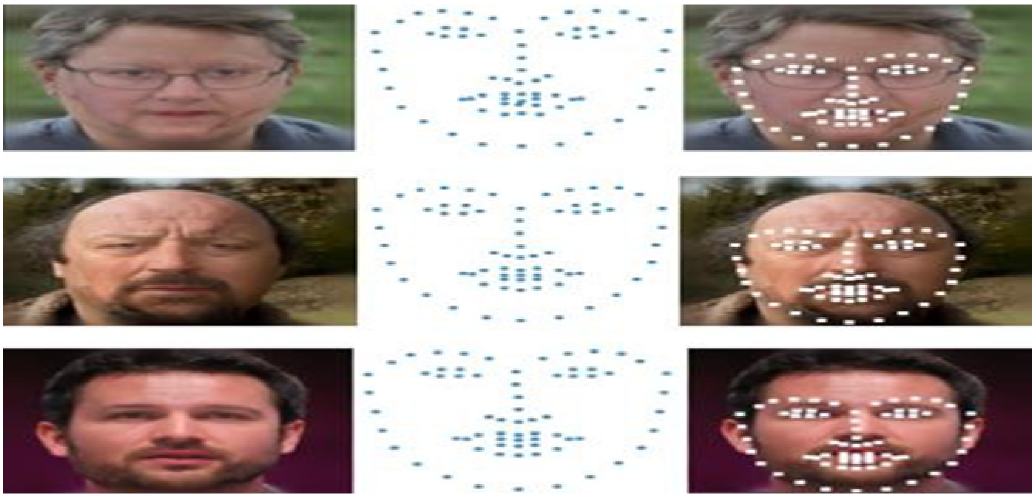

Figure 3. Input and output examples for landmark detection

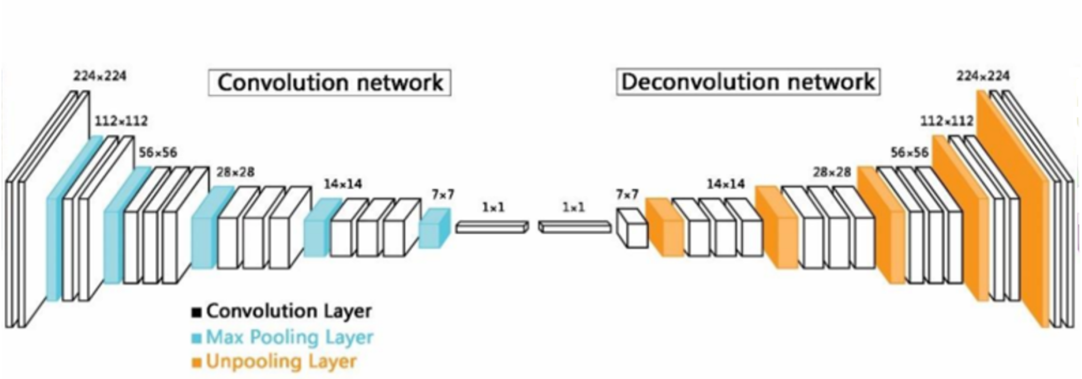

Figure 4. VGG-Face Model Convolution and Deconvolution network architecture

simply stacked together, but adds intermediate convolutional layers between them to literally invert the CNN not just functionally but also in architecture.

One more feature is that some intermediate layers can be removed to speed up the computations, which are represented by dashed lines in Figure 5. Finally, the first dense layer is responsible for changing the input vector to 4116-D to reach the desired output. A reshaping layer also exists after the dense layers, and for the last convolutional layer, the kernel size is one to reach an output having three channels with the desired dimensions. The convolutional model was used to predict the face texture. In training, the input to the model is a 4096-D feature vector coming from the encoded face using the VGG in testing, the input feature vector comes from the output of the audio encoder. The output matrix has the dimensions 224x224x3 and represent an RGB (Red, Green, Blue) image with the predicted texture. The model uses the conv transpose layer along with the normal conv layers to work as an inverter to the VGG network used for face encoding.

The model uses batch normalization layers after each block to speed up training by keeping the mean of the output around zero and the variance around one. The main advantage of using a dedicated model for predicting the texture of the face is that the model focuses on the shape of the face features and gives no concern to the positions of these features, making the required work needed by the model to minimize the loss less than using a decoder for the whole image. Another advantage of separating the face into texture and landmarks is robustness to changes and maintaining a good quality of the predicted face with changing the voice features coming from the audio encoder. The output texture of this model is warped with the output from the landmark model to produce the complete predicted face. The basic idea of image warping is to change the positions of the landmarks of the texture image to the positions of the landmarks coming from the landmarks model, making the predicted image obtain the predicted landmarks to form the final face prediction.

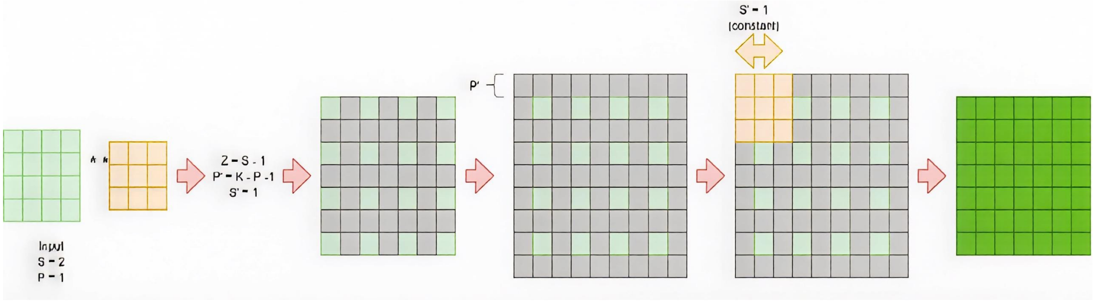

FIGURE 5. The operation of the Conv transpose layer

The texture model has a total number of parameters of 194,774,511. The total number of trainable parameters is 60,510,983. The total number of non-trainable parameters is 134,263,528. Figure 6 shows sample results of the texture model.

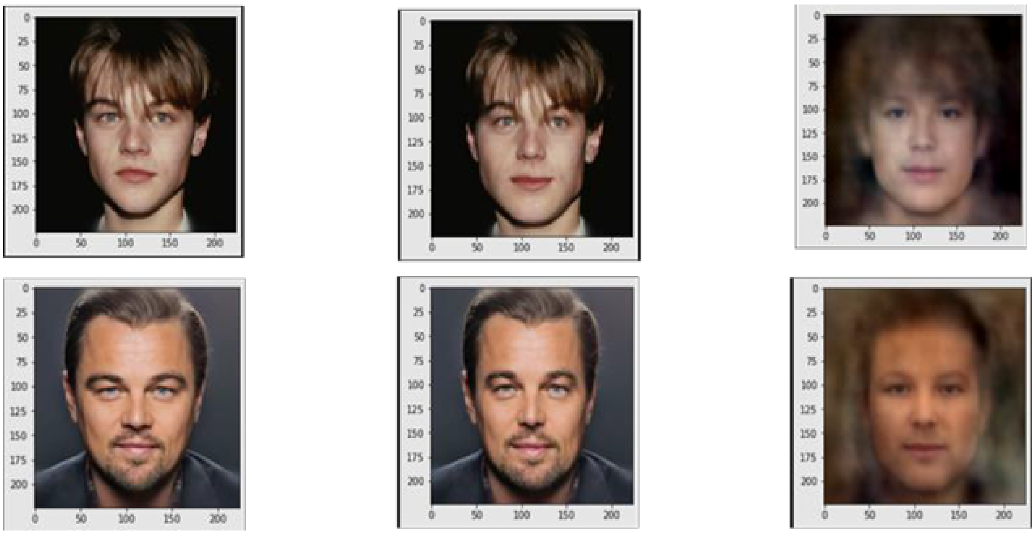

FIGURE 6. Sample results of the texture model. From left to right: input images, texture images, and predicted texture images.

4.5. **Loss Function.** Landmark model loss function. The landmarks model predicts the facial landmark positions from a given 4096-D facial feature vector. The landmarks are in the shape of two-element vectors representing the places of the facial landmarks (places of eyes, nose, and mouth).

The nature of the landmarks considers two types of losses between the ground truth landmarks and the predicted landmarks: mean squared error and mean absolute error. Equation (1) describes the mean squared error as:

$$MSE \quad = \quad \frac{1}{N}\sum_{i=1}^{N}\left(A_p^i - A_g^i\right)^2 \tag{1}$$

Where A_p is the flattened vector of predicted landmarks of dimensions (68*2 = 136), A_g is the flattened vector of ground truth landmarks of dimensions (68*2 = 136), and N is the number of entries in either of the flattened landmark vectors (N = 136). Equation (2) describes the mean absolute error as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left|T_p^i - T_g^i\right| \tag{2}$$

Comparing (1) and (2), we understand that (1) penalizes the high differences (more than one in value) more than the low differences because of the square function, which grows exponentially for large values of input, while (2) penalizes the high and low differences equally.

In the case of the landmarks model, when the differences between the ground truth landmarks and the predicted landmarks grow, the predicted landmarks will be in very visually unacceptable positions (for example, if an eye is slightly dispositioned, the result will be unpleasant). For further clarification, this is an example where we added a concentrated error of value 40 in only one point where the mean squared error (MSE) was 11.76, while the mean absolute error (MAE) was only 0.29. We added a distributed error of value 1360 where the MSE was 100 and the MAE was 10. Therefore, we chose the mean squared error to highly penalize the differences when exceeding one in value (concentrated loss); hence, it is more sensitive to outliers. Figure 7 demonstrates how the outliers were detected using the mean squared error.

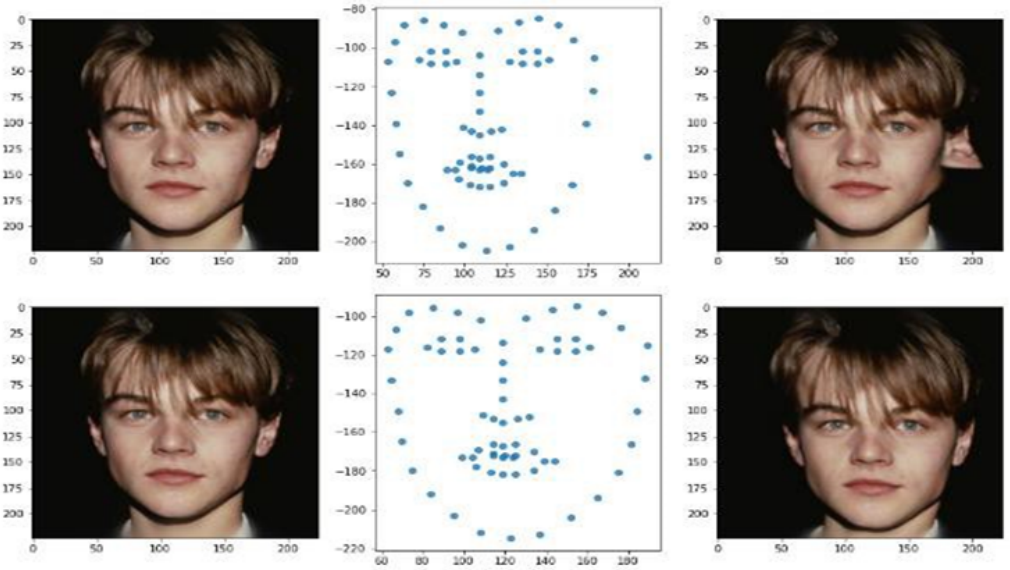

Figure 7. Clarification on using the mean squared error for the landmarks model loss function

4.6. **Texture model loss function.** The texture model predicts the texture of an image from a given 4096-D facial feature vector. The texture is mainly an image of the face, but its landmarks are moved to generic positions so that the textures have the same landmarks. Here, high differences between predicted texture image and ground truth texture image are not needed to be penalized more than small differences, such as in the landmark model. Therefore, we chose the loss to be the mean absolute error described by (2).

4.7. **Warping.** Warping is the process of manipulating an image based on certain criteria to distort it certainly. It changes the image shape using the geometric transformation as shown in Figure 8. Only the position of the points changes, but their intensity remains the same. In our application, we want to change the image such that the current landmarks are repositioned into the destination landmarks.

Transformation has many forms due to the nature of our application; therefore, we decided to use the differentiable spline interpolation as in equation (3):

$$f(x,y) = \sum_{i=1}^{N} w_i \, \varphi_k \left(\|(x,y) - (x_i, y_i)\| \right) + v_1 x + v_2 y + v_3 \tag{3}$$

Where φ_k is a radial basis function (RBF), whose value depends on the distance between the point and the center. RBF has many variations, but we used simply $\varphi_k = r$The parameters w, v_1, v_2, v_3 are chosen to satisfy the following constraints in equation(4) and equation(5):

$$\phi w + P v = s \tag{4}$$

$$\frac{N}{i-1} \sum w_i P_i = 0 \tag{5}$$

where P is the polynomial $v_1 x + v_2 y + v_3$.

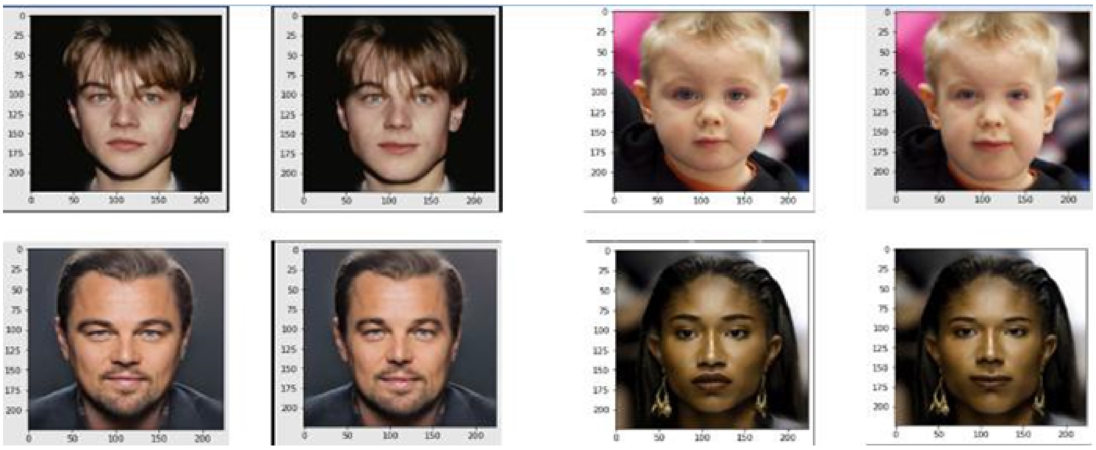

Figure 8. An example of image warping

We solve (4) and (5) together to obtain the parameters w, v_1, v_2, v_3 then we substitute in (4) to obtain the destination image as in Equation(6).

$$\begin{pmatrix} \phi & P \\ P^\top & 0 \end{pmatrix} \begin{pmatrix} \omega \\ v \end{pmatrix} = \begin{pmatrix} s \\ 0 \end{pmatrix} \tag{6}$$

5. **Training Data.**

5.1. **Data Collection.**

Data collection for the face decoder was a real issue for this project, not because of its lack, but because of the strict criteria needed for training the face decoder. A wide range of datasets were used for this task, some of which are:

1. The Chicago Face Database provides high-resolution, standardized photographs of male and female faces of varying ethnicity between the ages of 17 and 65 years [24]. 2. CelebAMask-HQ is a large-scale face image dataset that contains 30,000 high-resolution face images selected from the CelebA dataset by following CelebA-HQ [8].

3. The Oslo Face Database consists of approximately 200 male and female faces of neutral expression with three gaze directions: left, center, and right [25].

4. Flickr-Faces-HQ (FFHQ) is a high-quality image dataset of human faces, created as a benchmark for generative adversarial networks (GAN) which consists of 70,000 high-quality PNG images at 1024×1024 resolution and contains considerable variation in terms of age, ethnicity, image background, and many other small datasets that are publicly available at Kaggle.

5.2. **Face decoder data filtering.**

For training the face decoder, the images had to follow strict criteria. The Google Cloud Vision API was used to remove monochrome and blurry images, faces with high emotion scores (images that return anything other than very unlikely for joy, anger, surprise and sorrow), headwear or eyeglasses, and tilt, pan, or roll angles beyond 5°. After filtering, we have approximately 3 K images that are used for data augmentation. Images with multiple faces were either excluded if the criteria were not satisfied, or if a face fit the criteria, it was kept and cropped in further processing. Some examples of these images are shown in Figure 9.

5.3. **Data Augmentation.**

After selecting the most suitable thousand images, face morphing (many images are synthesized from a finite set of images using interpolation between image pairs) is applied to generate two thousand images for training the face decoder, as shown in Figure 10. These thousand images were hand-picked to ensure that they all strictly fit the criteria and had enough diversity in terms of race, age, and gender.

Our image dataset synthesizing model, illustrated in Figure 11, is a pipeline of operations applied sequentially. In short, two random images from the dataset are blended to generate a new image that can be considered an interpolation of both images.

First, random image A is selected randomly from 1000 images in the dataset. Then, a number of images represent the nearest neighbors of image A. We found that the closer the input images are to each other, the better the results obtained. Again, a random image B is selected from the 200 nearest neighbor images, and both images A and B are the source images of the synthesized image.

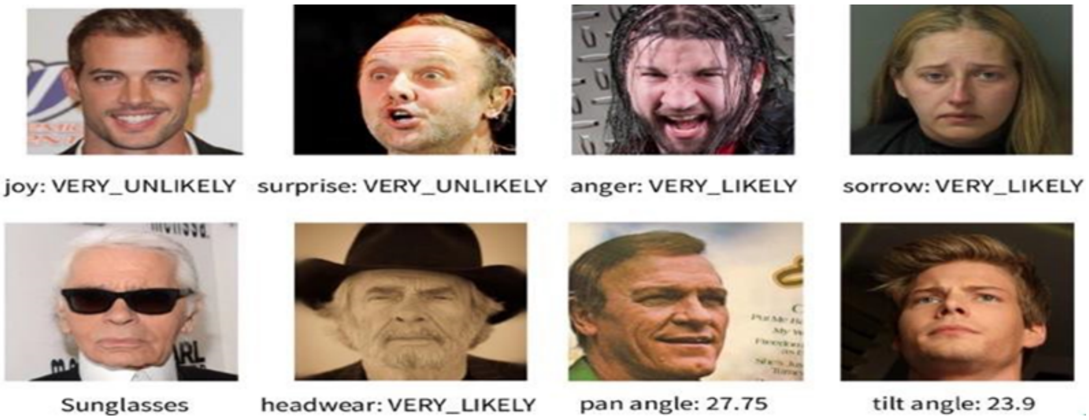

FIGURE 9. The sample of images that did not fit the required criteria.

FIGURE 10. The sample of input images for data augmentation.

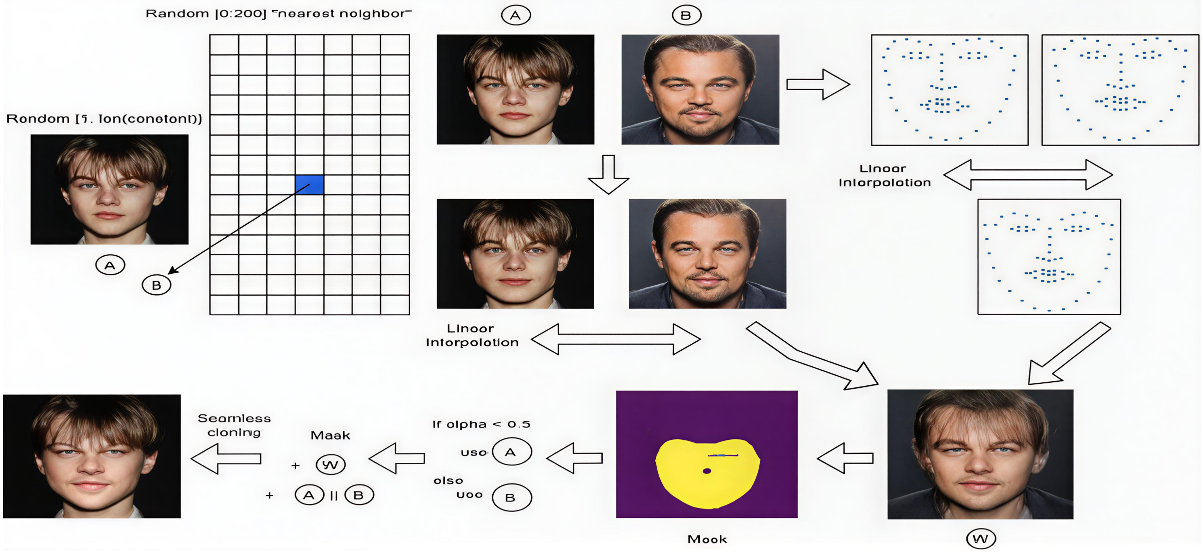

FIGURE 11. The face decoder dataset synthesizing model

Second, both image texture and landmarks are extracted for images A and B. Then, both the average texture and average landmarks generated in Figure 12 are found as inputs to the warping process, where the average texture is matched with the average landmarks.

Third, the image warping process is carried out to generate an image W of the averaged texture applied to the averaged face landmarks of the parent images. The point here is that the resulting image is hazy, especially in the area of hair and around the face boundary, as heads cannot be identical; therefore, we need a blending method to fix this issue.

Fourth, image enhancement is carried out on the warped image by automatically generating a mask whose boundary is obtained by joining the outer facial landmarks together, as explained in Figure 13. The mask is used to make one image beat the cuff of the other so that its facial features are more obvious.

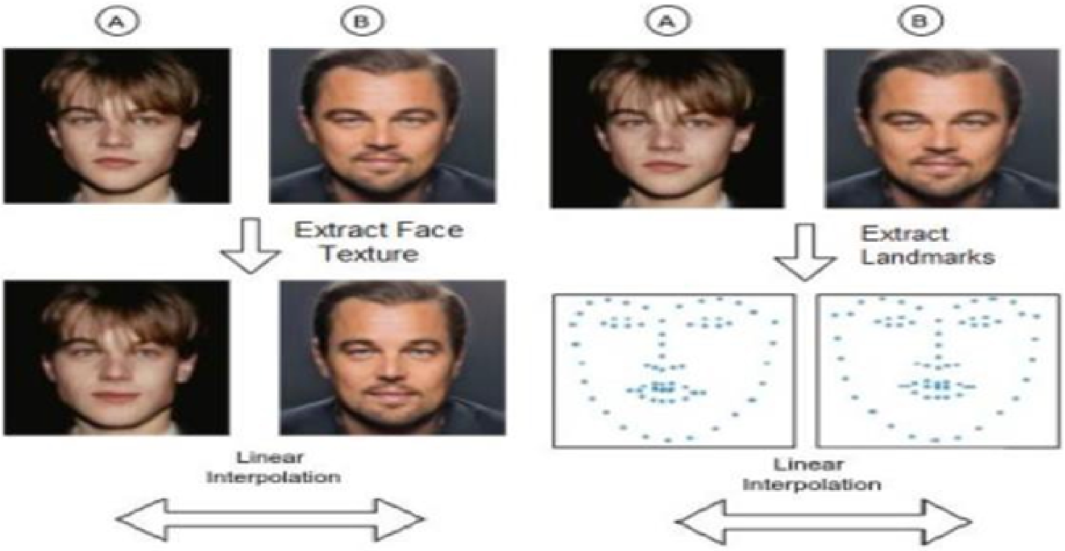

FIGURE 12. Generating average face texture and landmarks

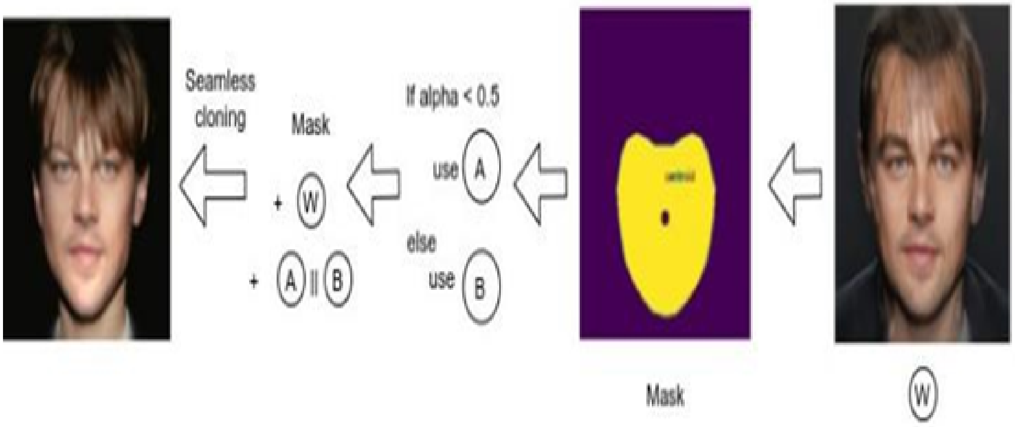

FIGURE 13. Generating and applying masks with seamless cloning.

The seamless cloning function is then introduced to generate the final and enhanced image.

OpenCV library provides a seamless cloning function whose inputs are source image, destination image, mask, center of the mask, and optional flags, returning an acceptable image that meets our criteria well, example output = cv2.seamlessClone(src, dst, mask, center, flags). The final images resulting from data augmentation fit the same criteria required for training the face decoder as shown in Figure 14.

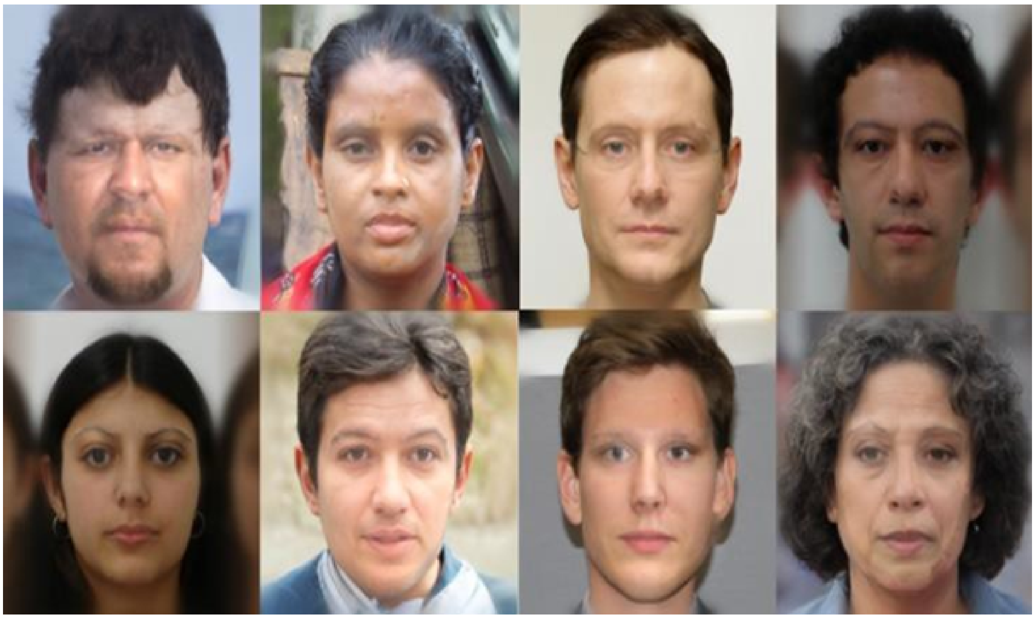

FIGURE 14. The sample of output images from data augmentation

6. **Experimental Results.** This section evaluates the performance of our proposed face reconstruction framework. We demonstrate the accuracy of the predicted landmarks, the fidelity of texture synthesis, and the realism of the final warped face reconstruction. The landmark model was trained on 200,000 examples over 200 epochs.

The proposed model is trained on a filtered and augmented dataset of frontal, neutral-expression faces, using VGG-Face embeddings as input. Evaluation metrics were computed on a held-out validation set. The landmark decoder's accuracy is measured using the Mean Squared Error (MSE) between predicted and ground truth 68-point coordinates. Table 1 shows the proposed model surpasses baselines in all metrics. It achieves higher SSIM and PSNR than FE-GAN, Speech2Face, and MaskGAN, indicating better fidelity. Its lower FID confirms more realistic faces compared to MaskGAN and FE-GAN. Landmark MSE is also significantly better, demonstrating precise geometric accuracy. These results validate separate landmark/texture training and robust warping, leading to visually coherent faces robust to pose/lighting.

Table 2 evaluates the impact of key components in the proposed architecture by comparing configurations through SSIM, PSNR, and FID metrics. The proposed model architecture achieves the best performance, underscoring the necessity of its multi-decoder design and warping mechanism. Removing critical components degrades results significantly. Single Decoder fails due to conflated learning, needing task-specific decoders. No Warping shows warping's geometric alignment importance. Improvements to the full Model prove decoupling/warping vital for fidelity, validating specialized modules/alignment for accuracy and realism.

TABLE 1. Comparison using Quantitative Metrics Against Baselines

Method	**SSIM**	**PSNR (dB)**	**FID**	**Landmark MSE**
FE-GAN [4]	0.82	25.1	45.7	29.8
Speech2Face [1]	0.78	23.4	52.3	35.2
MaskGAN [5]	0.85	26.7	38.9	27.1
proposed Model	**0.89**	**28.6**	**32.1**	**20.4**

The curves in Figure 15 show that the landmark decoder learns very quickly and stably. At the very start (epoch 1), both the training and validation MSEs exceed 400, reflecting untrained weights. However, by epoch 10, the loss has already dropped below 100. Between epochs 10 and 50, the curves still exhibit occasional spikes, more pronounced on validation, likely caused by mini-batch variability or learning rate adjustments; however, even these peaks remain well under 200 MSE. After epoch 50, both curves settle into a smooth downward trend, converging around 20 MSE by the final epochs. Importantly, the training

Table 2. Proposed architecture Enhancement Evaluation

Configuration	SSIM	PSNR	FID
Single Decoder	0.74	22.1	48.9
Without Warping	0.81	25.3	41.2
Joint Landmark-Texture Loss	0.85	26.8	35.7
Full Model	**0.89**	**28.6**	**32.1**

and validation losses remain tightly coupled throughout, indicating that the model is neither overfitting nor underfitting and generalizes the landmark predictions effectively across unseen data.

The curves in Figure 16 show the descending trend of the loss and validation loss with a final value of approximately 20 MSE. We showed only the last 180 epochs in the graph because the first 20 epochs were unstable. Separately, the texture model was trained on 200,000 examples for 40 epochs.

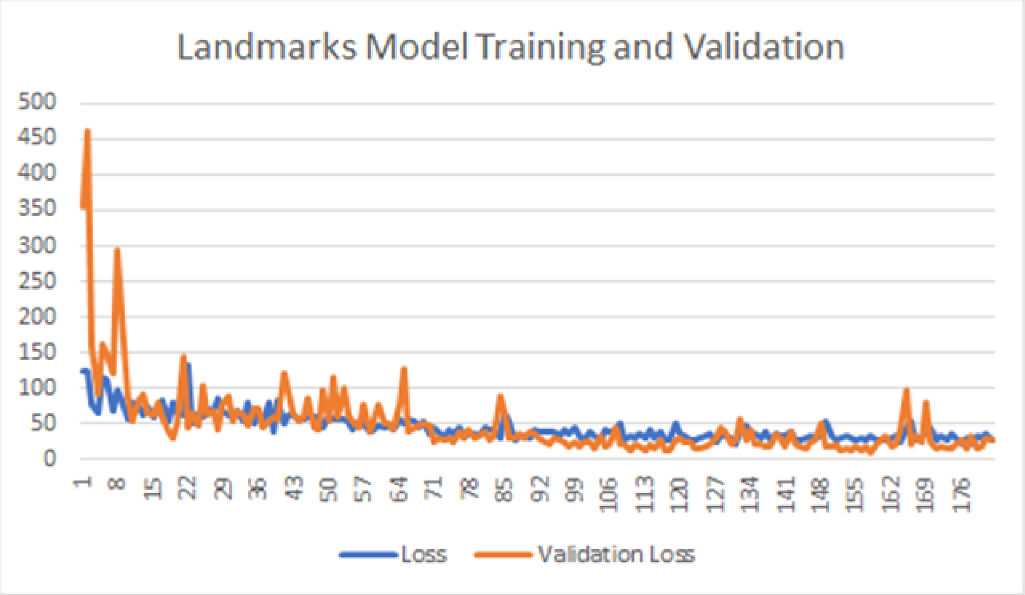

Figure 15. Landmark model training and validation

Figure 17 compares real faces (top) to model-generated faces (bottom). The model reconstructs faces well, capturing eye shape, nose, and lips, and handles varying lighting/pose, matching landmark positions. Fine textures are smoothed due to MAE loss. Age/gender are well captured, showing the model decodes biometrics effectively. Separating decoders works well, but texture could improve with adversarial training or better inputs.

The graph in Figure 18 demonstrates that the proposed model achieves a significantly lower FID score than FE-GAN, MaskGAN, and Speech2Face, indicating superior realism in generated faces. This likely results from its multi-decoder architecture, which improves reconstruction fidelity. The model's design and training minimize discrepancies, validating its ability to decode biometric correlations while maintaining visual coherence.

The Landmark Localization Error Heatmap in Figure 19 shows the proposed model efficiency. High accuracy (MSE¡10) for key facial landmarks (eyes, nose, lips) ensures reliable face reconstruction, validating the dedicated landmark decoder's effectiveness.

Higher errors in variable regions (jawline, outer lips; $MSE \approx 15 - 25$) demonstrate robustness to pose/expression changes without affecting core accuracy. Usability remains for applications like avatar creation, 95% of landmarks have$MSE < 30$, showing prioritized accuracy in critical areas. This modular design is more efficient than holistic approaches. Identified weaknesses (e.g., jawline) allow targeted improvements, highlighting the framework's interpretable design. Figure 19 confirms that the model efficiently balances accuracy and robustness, excelling in critical facial regions while gracefully handling variability in others by isolating and quantifying errors.

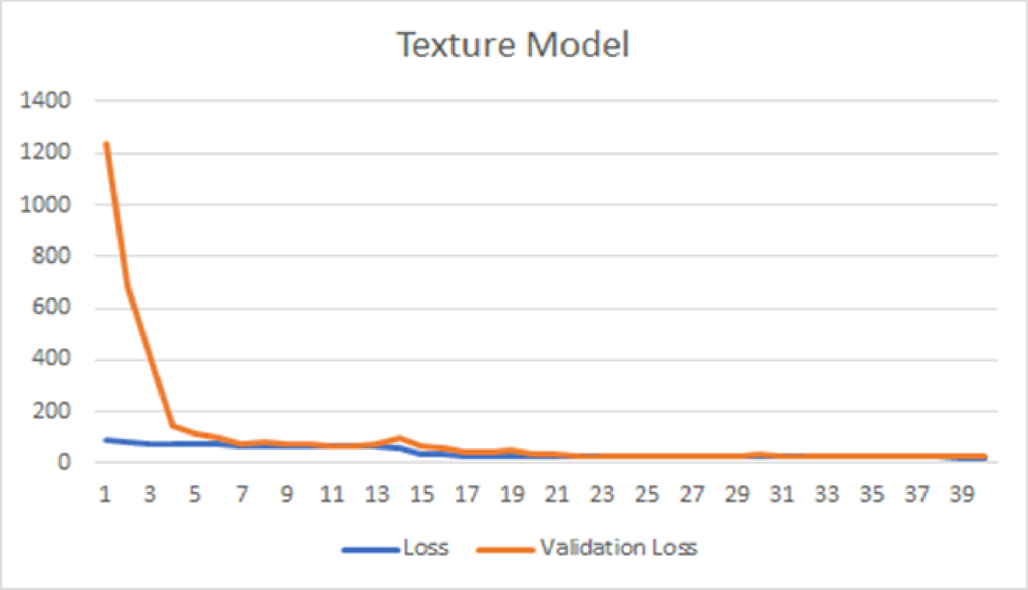

FIGURE 16. Texture model training and validation

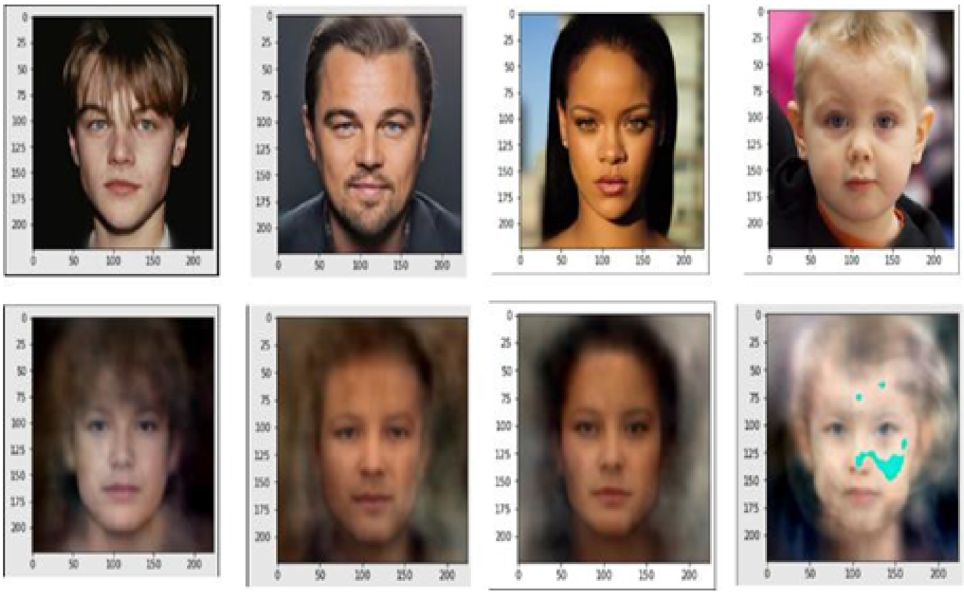

FIGURE 17. Sample visual results of the face decoder. (Top: Ground truth reference images. Bottom: Predicted images)

7. **Conclusion.** In this paper, we introduce the idea of how the facial structure affects various biometric features and how the face can be reconstructed by decoding the feature vector from the encoder of these biometric features, based on the correlation between them. Upon inspecting the results, we observed that the model can differentiate between the biometrics of individuals of different genders, ages, and races. Although the faces are different for different biometric inputs, they are not accurately representative of the real faces. We conclude that this method obtains the main features of the face but not the fine details.

8. **Limitation and Future Work.** Despite the efficiency of the proposed model, it has some limitations. While the model utilizes VGG-Face embeddings, it may potentially discard fine-grained texture features, leading to more uniform reconstructions. Also , the training data is limited to neutral, frontal faces; hence, severe emotions, occlusions, or significant position fluctuations are not well addressed. Ultimately, landmark prediction inaccuracies in peripheral areas may spread through the warping phase, somewhat influencing border realism.

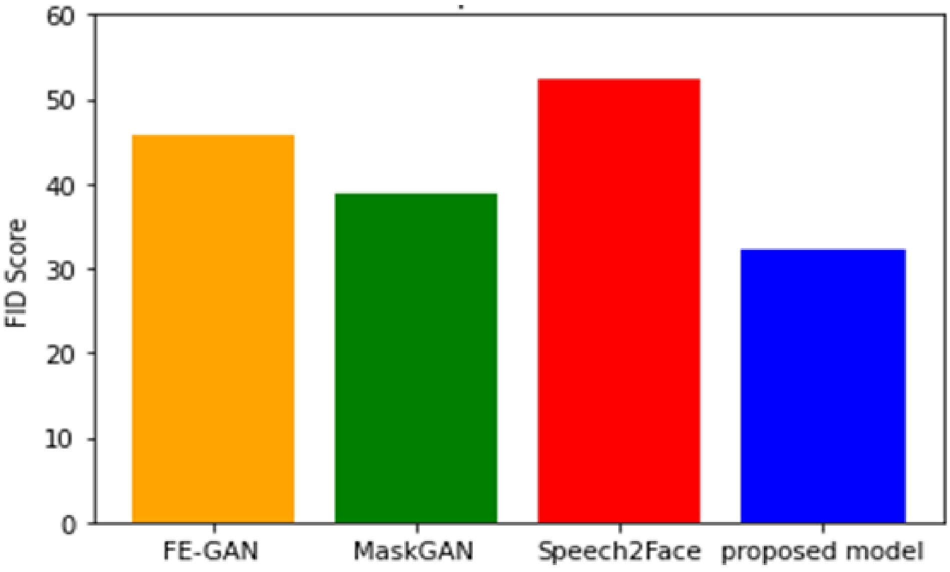

Figure 18. FID Score Comparison with baseline Models.

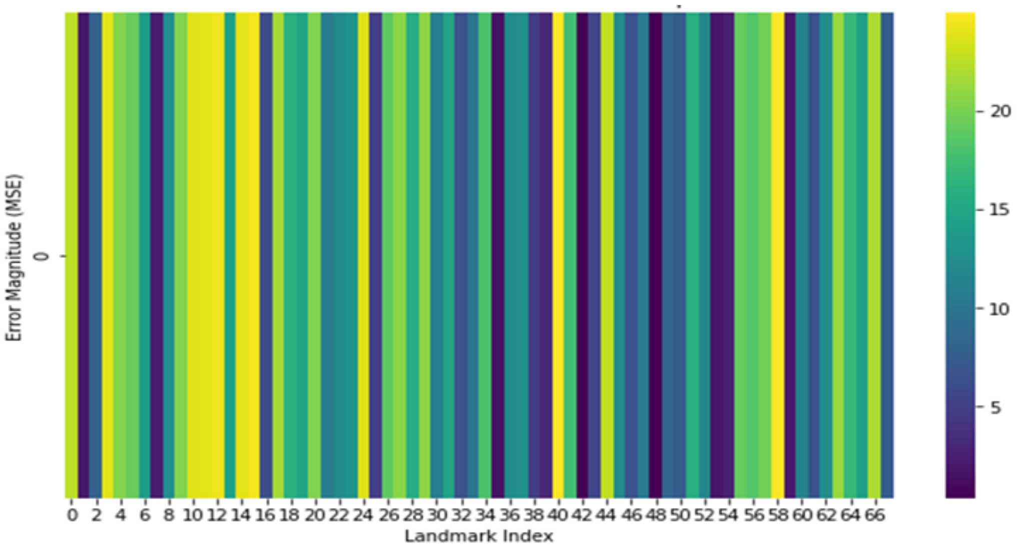

Figure 19. Landmark Localization Error Heatmap

In future work, we recommend increasing the biometric feature data length that is used for the extraction of the output embedded feature vector, e.g., the length of the voice recordings for speech-to-face reconstruction, which will also increase the computations and training time, but will probably improve the biometric encoder. In addition, choosing real (not generated through augmentation or GANs) neutral frontal-facing faces for training the face decoder may improve the model. Improving the architecture of the models and increasing the training time. Taking it a step further, one might try to reconstruct moving frames, not images, or even recreate facial expressions.

Acknowledgment. The author is grateful to the editorial staff and reviewers, as well as the corresponding author, who offered assistance in the form of advice, assessment, and checking during the study period, which has improved the presentation.

REFERENCES

[1] T.-H. Oh, T. Dekel, C. Kim, I. Mosseri, W. T. Freeman *et al.*, "Speech2Face: Learning the face behind a voice," in *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, ”Synthesizing normalized faces from facial identity features,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Z. M. Chan, C. Y. Lau, and K. F. Thang, ”Visual speech recognition of lips images using convolutional neural network in VGG-M model,” *Journal of Information Hiding and Multimedia Signal Processing*,vol. 11, no. 3, 2020.

[4] A. S. Aziz, H. K. Mohamed, and A. Abdelhafeez, ”Unveiling the power of convolutional networks: Applied computational intelligence for arrhythmia detection from ECG signals,” *International Journal of Advances in Applied Computational Intelligence*, vol. 1, no. 2, 2022, doi: 10.54216/ijaaci.010205.

[5] N. Datta, J. Sikder, R. Chakma, and R. K. Das, ”Head features-based deep learning approach for recognizing emotion, gender and age,” *Journal of Information Hiding and Multimedia Signal Processing*, vol. 14, no. 4, 2023.

[6] K. Balamurali, S. Chandru, M. S. Razvi, and V. Sathiesh Kumar, ”Face spoof detection using VGG-Face architecture,” *Journal of Physics: Conference Series*, 2021, doi: 10.1088/1742-6596/1917/1/012010.

[7] Z. Fang, Z. Liu, T. Liu, C.-C. Hung, J. Xiao, and G. Feng, ”Facial expression GAN for voice-driven face generation,” *The Visual Computer*, vol. 38, no. 3, 2022, doi: 10.1007/s00371-021-02074-w.

[8] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, ”MaskGAN: Towards diverse and interactive facial image manipulation,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*,2020.

[9] Y. Wu and Q. Ji, ”Facial landmark detection: A literature survey,” *Int. J. of Computer Vision*, vol. 127, no. 2, pp. 115–142, 2019.

[10] F. Cole, D. Belanger, D. Krishnan, A. Sarna, I. Mosseri, and W. T. Freeman, ”Synthesizing normalized faces from facial identity features,” *Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, doi: 10.1109/CVPR.2017.361.

[11] M. Li, W. A. P. Smith, and P. Huber, ”ID2image: Leakage of Non-ID Information into Face Descriptors and Inversion from Descriptors to Images,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2023, doi: 10.1007/978-3-031-31438-4_29.

[12] J. Križaj, R. O. Plesh, M. Banavar, S. Schuckers, and V. Štruc, ”Deep Face Decoder: Towards understanding the embedding space of convolutional networks through visual reconstruction of deep face templates,” *Engineering Applications of Artificial Intelligence*, vol. 132, 2024, doi: 10.1016/j.engappai.2024.107941.

[13] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, ”Synthesizing Obama: Learning lip sync from audio,” *ACM Trans. Graph. (SIGGRAPH)*, vol. 36, no. 4, 2017.

[14] N. Sadoughi and C. Busso, ”Speech-driven expressive talking lips with conditional sequential generative adversarial networks,” *IEEE Trans. on Affective Computing*, 2019.

[15] S. L. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. K. Hodgins, and I. A. Matthews, ”A deep learning approach for generalized speech animation,” *ACM Trans. Graph. (SIGGRAPH)*, vol. 36, no. 4, pp. 93:1–93:11, 2017.

[16] T. Karras *et al.*, ”Audio-driven facial animation by joint end-to-end learning of pose and emotion,” *ACM Trans. Graph. (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[17] O. Wiles, A. Koepke, and A. Zisserman, ”X2Face: A network for controlling face generation using images, audio, and pose codes,” in *Proc. European Conf. on Computer Vision (ECCV)*, 2018.

[18] X. Li, X. Li, and J. Deng, ”Disentangled representation transformer network for 3D face reconstruction and robust dense alignment,” *Visual Computer*, vol. 40, no. 11, 2024, doi: 10.1007/s00371-023-03202-4.

[19] J. Liang, H. Liu, H. Xu, and D. Luo, ”Generalizable Face Landmarking Guided by Conditional Face Warping,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2024, doi: 10.1109/CVPR52733.2024.00235.

[20] O. M. Parkhi, A. Vedaldi, and A. Zisserman, ”Deep Face Recognition,” *British Machine Vision Conference*, 2015.

[21] I. A. Kakadiaris *et al.*, ”Profile-based face recognition,” in *Proc. 8th IEEE Int. Conf. on Automatic Face & Gesture Recognition*, 2008.

[22] D. P. Kingma and J. Ba, ”Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[23] E. Maggiori *et al.*, ”High-resolution aerial image labeling with convolutional neural networks,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7092–7103, 2017.

[24] D. S. Ma, J. Correll, and B. Wittenbrink, ”The Chicago face database: A free stimulus set of faces and norming data,” *Behavior Research Methods*, vol. 47, pp. 1122–1135, 2015.

[25] O. Chelnokova *et al.*, ”Rewards of beauty: The opioid system mediates social motivation in humans,” *Molecular Psychiatry*, vol. 19, no. 7, pp. 746–747, 2014.