Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1, 131-145
2025
Publisher: Learning Gate
DOI: 10.55214/25768484.v9i1.3883

VN3DFace: 3D facial expression database of Vietnamese

Thi Chau Ma[1*], Dang Ha Nguyen[2]
[1,2]University of Engineering and Technology, Hanoi, Vietnam; chaumt@vnu.edu.vn (T.C.M.) ndhauetvnu@gmail.com (D.H.N.).

Abstract: The availability of 3D face datasets has surged, enhancing computer vision technologies and enabling better facial recognition systems. However, unique facial features across different ethnic groups necessitate tailored datasets. Despite their importance, creating 3D facial data is resource-intensive, requiring advanced hardware and expertise. This paper presents a solution using 3D face reconstruction techniques to develop a 3D emotional face dataset from images of Vietnamese individuals, named VN3DFace. This approach can be applied to any 2D facial image dataset, facilitating the generation of larger, diverse datasets. VN3DFace demonstrates promising results, achieving an LPIPS score of 0.09 and a DFID score of 0.25, showcasing its potential for various applications.

Keywords: *3D facial reconstruction, 3D facial database, Facial expression.*

1. Introduction

In contemporary research across diverse fields such as computer vision, biometrics, facial recognition, and virtual reality, the importance of 3D face model datasets cannot be overstated. These datasets play a crucial role in the development and evaluation of algorithms, training machine learning models, and advancing our understanding of facial anatomy and expressions. One of the primary benefits of 3D face models is their ability to capture the intricate three-dimensional geometry of the human face, providing a more comprehensive representation compared to traditional 2D images. This enhanced representation contributes to the precision and robustness of facial recognition systems, enabling more accurate identification and authentication processes, particularly in security applications. Additionally, 3D faces datasets make significant contributions to the creation of realistic avatars in virtual environments. They facilitate immersive experiences in gaming, simulations, and virtual communication by providing detailed facial structures and expressions.

Previously, research groups had to rely on complex systems to accurately capture or scan subjects and generate realistic 3D face models. However, the creation of large datasets using these methods posed challenges due to factors such as expensive hardware, lengthy capturing processes, and the need

Figure 1.
2D face images.

History: Received: 7 November 2024; Revised: 5 December 2024; Accepted: 26 December 2024; Published: 6 January 2025
* Correspondence: chaumt@vnu.edu.vn

For physical presence of subjects [1-3]. To overcome these obstacles, we have developed 3D face reconstruction frameworks aimed at creating a dataset of Vietnamese individuals without encountering these limitations. Our approach involves leveraging DECA and MICA technologies [4, 5] to extract features from 2D images (Figure 1) and modifying the FLAME model [6] to transform the base model into desired 3D faces (Figure 2). Additionally, we employ a specialized model for learning and transferring facial expressions from the input image to the 3D face model. To evaluate the effectiveness of our method, we conducted experiments using the FaceScape dataset in various evaluation scenarios. Our contributions include:

- Introducing an emotionally aggressive model that combines with other state-of-the-art 3D face reconstruction methods to accurately capture the intricacies of the human face.
- Creating a new and unique 3D face database named" VN3D face," specifically designed to cater to the facial characteristics of Vietnamese individuals.

Figure 2.
3D face models.

The paper is organized as follows Session 2 introduces the related works. The process of constructing our 3DVNFace database is described in Session 3. In Session 4, we present an application of the 3DVNFace database to the emotion retargeting problem. Session 5 outlines the methods and scenarios for qualitative and quantitative evaluation of the constructed database. The conclusion is presented in Session 6.

2. Related Works

2.1. 3D Face Dataset

The concept of 3D Morphable Models (3DMMs) was initially introduced by Blanz and Vetter [7]. These models utilize statistical techniques such as Principal Component Analysis (PCA) to capture variations in facial shape and texture. By doing so, they can represent a wide range of human faces using a small set of parameters. Numerous studies have been conducted to develop generalized 3DMMs.

One prominent solution for 3D face modeling is the Faces Learned with an Articulated Model and Expressions (FLAME) [6]. FLAME is a 3DMM model capable of generating realistic and expressive 3D face models that accurately capture variations in facial shape and expression. It combines a linear shape space trained with thousands of human head scans with an articulated jaw, neck, eyeballs, pose dependent corrective blendshapes, and additional global expression blendshapes. The model learns articulations based on pose and expression from 4D face sequences found in the D3DFACS dataset and other sources. FLAME separates the representation of identity, pose, and facial expression into distinct parameter spaces and combines them using linear blend skinning (LBS) and blendshapes. Its ability to generate highly expressive 3D face models makes it a fundamental component in many advanced face reconstruction models.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

FaceWarehouse [8] is a database specifically designed for visual computing applications, with a focus on 3D facial expressions. This database demonstrates its potential through various applications, including facial image manipulation, face component transfer, real-time performance-based facial image animation, and converting facial animation from video to image. FaceWarehouse was created by capturing data from a diverse group of 150 individuals using the Kinect RGBD camera. Each person had multiple expressions recorded, allowing for a comprehensive collection of matched expressions. The database provides a broad range of facial actions for everyone, enabling the depiction of various human facial expressions.

Wood, et al. [9] propose a novel method that streamlines the face reconstruction process by leveraging an increased number of landmarks. The study shows that dense landmarks are effective in integrating facial shape information across frames, resulting in precise and expressive capture of facial performance in both monocular and multi-view scenarios. The method can accurately predict dense landmarks and fit 3D face models at a high speed, even surpassing 150 frames per second on a single CPU thread. By accurately predicting a larger number of landmarks, including crucial areas like the eyes and teeth, the method achieves state-of-the-art results in monocular 3D face reconstruction.

The FaceScape dataset [10] showcases a detailed collection of 3D face models. It introduces an algorithm capable of generating intricate and poseable 3D face models from a single image input. The dataset consists of thousands of textured 3D faces obtained from hundreds of individuals, covering a wide range of expressions. These models provide a detailed representation of facial geometry at a fine level, ensuring topological uniformity.

These detailed 3D facial models are represented using a combination of a 3D morphable model for general shapes and displacement maps for precise geometry. A pioneering algorithm is proposed that utilizes a deep neural network to learn expression-specific dynamic details from the FaceScape dataset. This algorithm forms the basis of a system that can predict 3D face models from a single image input, distinguishing it from previous methods.

2.2. Single Image-Based 3D Face Reconstruction

Detailed Expression Capture and Animation (DECA) [4] is a method that utilizes FLAME as a key component for reconstructing 3D face models from single-view images. DECA not only detects facial shape and expression but also accurately maps facial texture from the image to the 3D model by leveraging a 3D texture space. During training, DECA achieves exceptional shape reconstruction performance by incorporating a shape consistency loss. Additionally, DECA introduces a novel detail consistency loss that separates expression-dependent wrinkles from person-specific details. The use of a low-dimensional detail latent space enhances the robustness of fine-scale reconstruction against noise and occlusions. The novel loss function facilitates the disentanglement of identity-related details from expression-dependent wrinkle details.

MetrIC fAce (MICA) [5] is another method that focuses on reconstructing 3D face models from single-view images and also incorporates FLAME as a component in the reconstruction process. However, there are notable differences between DECA and MICA. Unlike DECA, MICA solely focuses on reconstructing the 3D face model using shape parameters β and does not involve the reconstruction of facial expressions or textures. MICA adopts a distinct approach for reconstructing the shape parameters by utilizing the ArcFace architecture as the basis for its network and refining the last three ResNet blocks of this architecture. The use of this refined network in MICA results in a more accurate reconstruction of the shape parameters compared to DECA, as demonstrated in the NoW benchmark.

The Basel Face Model (BFM) is a morphable face model initially introduced by Paysan et al. in 2009. Over time, the University of Basel has released three official versions of the BFM: 2009, 2017, and 2019. In the field of 3D face reconstruction, Deng et al. have introduced a highly effective method called Deep3D [11]. This method enables the reconstruction of a person's 3D face model using both single and multiple images. To overcome the absence of ground truth 3D data, Deng et al. devised a new hybrid robust loss function that incorporates both low-level and perception-level information. They

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

employed this modified loss function to train a ResNet50 network, with the number of neurons in the fully connected layer adjusted to 254, corresponding to the required coefficients. These coefficients are then combined with the BFM, a generic face model, to generate the desired human face.

In another study Paier, et al. [12] the authors introduce a novel hybrid animation framework that leverages recent advancements in deep learning to create an interactive animation engine for editing facial expressions. Despite the progress made in computer graphics, generating realistic animations of human faces remains a challenging task. To address these challenges, the authors propose an automatic pipeline for generating training sequences that combine dynamic textures with consistent three-dimensional face models. They use this dataset to train a variational autoencoder, which learns a low-dimensional latent space of facial expressions. This latent space is then employed for interactive facial animation, enabling users to edit and manipulate facial expressions.

Yenamandra, et al. [13] introduce a novel deep implicit 3D morphable model (i3DMM) capable of representing full heads, including hair, while capturing identity-specific geometry, texture, and expressions of the frontal face. To train the i3DMM, they collect a new dataset comprising 64 individuals with various expressions and hairstyles. The proposed approach offers several advantages: It is the first morphable model that encompasses the entire head, can be trained on rigidly aligned scans without the need for challenging non-rigid registration, decouples the shape model into an implicit reference shape and its deformations, enabling the learning of dense correspondences between shapes implicitly, and facilitates the semantic disentanglement of geometry and color components. The authors demonstrate the effectiveness of the i3DMM through ablation studies and comparisons with state-of-the-art models.

In their research Zhuang, et al. [14] the authors propose a parametric model called Morphable Facial NeRF (MoFaNeRF), which utilizes a neural radiance field to map free-view images into a vector space representing coded facial shape, expression, and appearance. The model takes coded facial attributes, spatial coordinates, and view direction as input to a multi-layer perceptron (MLP) and generates the radiance of each point in space for photo-realistic image synthesis. MoFaNeRF excels in synthesizing highly realistic facial details, including those of the eyes, mouth, and beard. Additionally, the model enables seamless face morphing by interpolating the input shape, expression, and appearance codes. Through the incorporation of identity-specific modulation and a texture encoder, the model accurately captures photometric details and exhibits strong representation capabilities. Experimental results show that the proposed method surpasses previous parametric models in terms of representation ability and achieves competitive performance in multiple applications.

2.3. Expression Transfer

The conventional metrics commonly used for training, such as landmark reprojection error, photometric error, and face recognition loss, have limitations in accurately capturing facial expressions with a high level of fidelity. These metrics fail to effectively convey the emotional nuances depicted in the input image, resulting in reconstructed facial geometries that do not fully capture the intended expressions. To address this limitation, a novel solution called EMOtion Capture and Animation (EMOCA) [15] has been proposed. During the training phase of EMOCA, a deep perceptual emotion consistency loss is introduced. This loss ensures that the reconstructed 3D expression aligns with the emotional expression portrayed in the input image.

As the use of 3D facial avatars becomes increasingly prevalent in communication, it is crucial that they can effectively convey emotions. Unfortunately, existing state-of-the-art methods that regress parametric 3D face models from monocular images struggle to capture the full spectrum of facial expressions, including subtle or extreme emotions. In contrast, EMOCA achieves comparable levels of 3D reconstruction errors as the current best methods while surpassing them in terms of the quality of the reconstructed expression and the perceived emotional content. Furthermore, the research extends to directly include the regression of valence and arousal levels, as well as the classification of basic expressions based on the estimated 3D face parameters. In the context of recognizing emotions in real-

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

world situations, EMOCA's purely geometric approach performs on par with the leading image-based methods, emphasizing the value of 3D geometry in the analysis of human behavior. Traditional methods used for image-based 3D face reconstruction and facial motion retargeting, which involve fitting a 3D morphable model (3DMM) to the face, have limitations in their modeling capacity and struggle to generalize well to diverse real-world data. Techniques such as deformation transfer or multilinear tensor fail to adequately address the variations in local and global skin deformations that occur in different individuals due to facial expressions. Additionally, these methods typically learn a single albedo (surface reflectance) per user, failing to capture expression-specific skin reflectance variations.

To overcome these limitations, the authors propose an end-to-end framework in Chaudhuri, et al. [16] that simultaneously learns a personalized face model for each user and per-frame facial motion parameters using a large dataset of in-the-wild videos capturing user expressions. This framework predicts personalized corrections on top of a 3DMM prior, allowing the learning of user-specific expression blendshapes and dynamic albedo maps that are specific to each expression. Novel constraints are introduced to ensure that the corrected blendshapes preserve their semantic meanings and that the reconstructed geometry is separated from the albedo. Experimental results demonstrate that this personalized approach accurately captures fine-grained facial dynamics across a wide range of conditions, resulting in more precise face reconstruction and facial motion retargeting compared to stateof-the-art methods.

In Moser, et al. [17] a simple algorithm is proposed for automatically transferring facial expressions from videos to a 3D character and between different 3D characters through their rendered animations. The method learns a shared latent representation that is semantically consistent across various input image domains using an unsupervised image-to-image translation model. It also learns linear mapping, in a supervised manner, from the encoded representation of character images to the animation coefficients. During inference, when provided with the source domain (i.e., actor footage), the algorithm predicts the corresponding animation coefficients for the target character, enabling automatic remapping of expressions between different identities, even when there are differences in physiognomy. The technique is demonstrated in the context of marker less motion capture with controlled lighting conditions, both for a single actor and multiple actors. Furthermore, it showcases the ability to automatically transfer facial animation between distinct characters without relying on consistent mesh parameterization or engineered geometric priors. The method is compared against standard approaches commonly used in production and recent state-of-the-art models in the field of single-camera face tracking.

3. VN3D Face Database Building

$T(I, M)$ is defined as a pair that consists of a facial image I and its corresponding 3D model M. The VN3DFace database comprises facial expression images and their 3D models of 400 Vietnamese individuals (subjects). Each subject is associated with 7 pairs of facial expression images and their respective 3D models, denoted as $Ti(Ii, Mi)$, with $i = 1..7$. The construction of VN3DFace involves two main steps: (i) collecting facial expression images, which involves capturing 7 facial expression images for each subject, and (ii) generating the corresponding facial expression 3D models from the captured images.

3.1. Collecting Facial Expression Images

We utilized a multi-camera system comprising 27 Full HD DSLR Canon EOS 1500D cameras and 10 lighting devices to capture images of each subject. The cameras are organized into three layers, with each layer featuring cameras evenly spaced apart (Figure 3). We employed 11 distinct lighting setups, adjusting intensities from 0 to 1,000 lux and varying lighting angles to achieve partial facial shadowing. It is ensured that the tip of each subject's nose is positioned centrally within the frame. Synchronization of all capturing devices ensures simultaneous image capture, streamlining the process. Each camera is set up with the following configurations (Table 1).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

Table 1.
Camera configuration.

Shutter speed	Aperture	ISO	White balance	Forcus	Image quality	Zoom
1/40s	F5.6	800	AWB	Manual focus	S1	24

Each participant undergoes a roughly one-hour procedure, divided into five brief sessions. The initial three sessions involve capturing facial images without any accessories. In the fourth and fifth sessions, participants are instructed to wear a face mask and sunglasses respectively. Data gathered from these final two sessions contributes to enhancing facial reconstruction in scenarios involving occlusions. During each session, individuals are required to display seven distinct expressions: neutral, happy, sad, afraid, angry, surprised, and disgusted, across 11 mentioned lighting setups. Consequently, the total number of images obtained from each individual amount to 10,395, calculated as 5 (capturing sessions) × 11 (lighting conditions) × 7 (facial expressions) × 27 (poses). The dataset comprises data obtained from subjects aged between 20 and 50, with a predominant representation in their twenties. The gender distribution is nearly equal, with 48% male and 52% female subjects. Additional information regarding the dataset is available in [18].

Figure 3.
Data capture system.

3.2. Generating Facial Expression 3D Models

To generate the 3D head from a single input image, we utilized an architecture that combines DECA [4] MICA [5] and our emotional regressive model. The architecture, as depicted in Figure 4, follows the following process.

Firstly, DECA takes the input image and estimates parameters such as camera (c), albedo (α), light (l), expression (ψ), identity's shape (β), head pose (θ), and detail (δ). However, for the FLAME model, three parameters—expression (ψ), identity's shape (β), and head pose (θ)—are replaced by the outputs of MICA and the emotional regressive model. Although DECA could be used for the identity's shape parameters, we opted for MICA due to its ability to reconstruct these parameters with metric accuracy and produce superior results, as evidenced in the NoW Challenge benchmarks.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

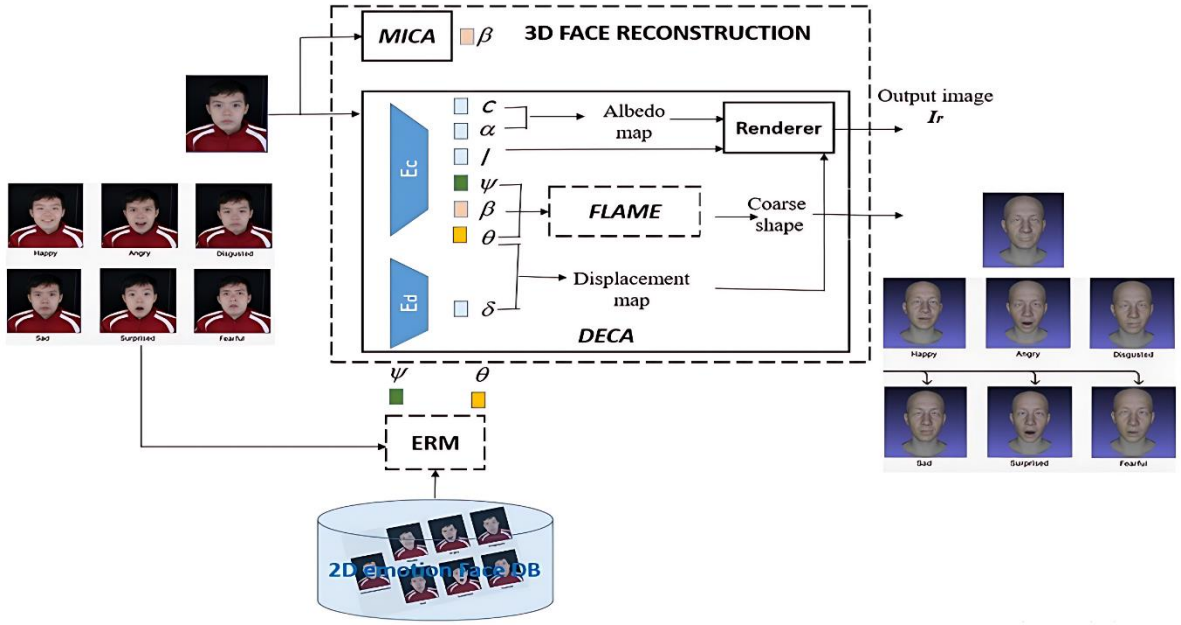

Figure 4.
Process of 3D expression reconstruction from images.

The two parameters, expression (ψ) and head pose (θ), are employed to generate a 3D facial model, with the expressions being replaced by the output of our emotional regression model (presented in section 3.2). The resulting FLAME mesh, along with the displacement map, is used to create a detailed 3D facial model. Additionally, the albedo map, generated from the camera and light parameters, is applied to the detailed 3D facial model, which is then rendered to produce an output image, Ir. This output image, Ir, is further utilized to evaluate the detailed 3D facial model.

3.3. Expression Learning

Ekman [19] the developer of a widely accepted theory on basic emotions and their expressions, proposes that there are six fundamental emotions: sadness, happiness, fear, anger, surprise, and disgust. Expression generation models significantly depend on datasets. The quality and diversity of these datasets influence the model's ability to learn and accurately reproduce expressions. A comprehensive dataset that includes a variety of expressions, lighting conditions, angles, and facial features will enhance the model's learning and enable it to generate more natural expressions. Furthermore, each ethnicity possesses unique and subtle characteristics in their facial features; therefore, if the dataset represents a specific demographic or a range of expressions, the model must be retrained on a dataset that accurately reflects that demographic and the desired expressions.

To generate expression parameters from images that can be seamlessly integrated during the reconstruction of 3D faces based on DECA/MICA, our idea is to create an emotional-regressive model to convert emotional labels to FLAME's pose and expression parameters. The emotional regressive model is trained on the V-Face dataset [18] with the ground truth acquired by running the dataset through the pre-trained face reconstruction model (DECA/ MICA). The Emotional Regressive Model Architecture (called ERM) is shown in Figure 5 below. The ERM model is a feed-forward model with 6 input nodes (for 6 expressions) and 56 output nodes (50 for expression parameters and 6 for pose parameters).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

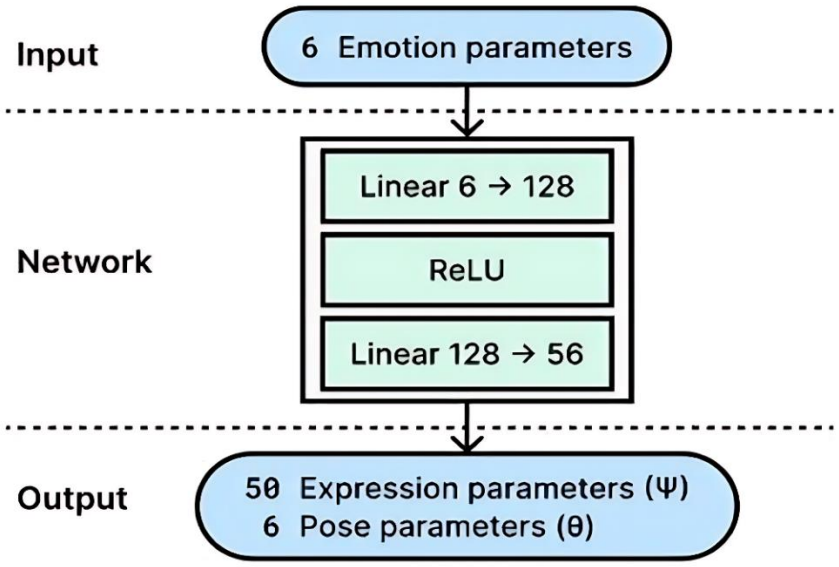

Figure 5.
The emotional regressive model.

The" linear" layers perform linear transformations on the input data, which means they apply a weight matrix and a bias vector to the input. The ReLU activation function is used to introduce nonlinearity into the network. Since the emotion parameters in FLAME are projected linearly to the mesh vertices, the linear layers should perform well as a linear regressor. The loss function used is the mean absolute error (MAE) loss function which is defined as:

$$L_{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (1)$$

*where y_i is the ground truth, $\hat{y}_i$ is the predicted value, and N is the number of samples. The loss function is d*efined as:

$$Loss = Loss(p_{exp}, \hat{P}_{exp}) + Loss(p_{pose}, \hat{P}_{pose}) \quad (2)$$

where:

$$Loss(p_{exp}, \hat{P}_{exp}) = L_{MAE}(p_{exp}, \hat{P}_{exp}) \quad (3)$$

and

$$Loss(p_{pose}, \hat{P}_{pose}) = \rho L_{MAE}(p_{neckpose}, \hat{P}_{neckpose}) + \epsilon L_{MAE}(p_{globalpose}, \hat{P}_{globalpose}) \quad (4)$$

and p_{exp} is the ground truth expression parameters, $\hat{P}_{exp}$ is the predicted expression parameters, P_{pose} is the ground truth pose parameters, $\hat{p}_{pose}$is the predicted pose parameters. Within pose parameters, there are neck pose parameters and global pose parameters. ρ and ϵ are the weights for the neck pose and global pose loss, respectively.

Following the training process, we successfully acquired the ERM model, which effectively learns expression parameters and pose parameters for the reconstruction of 3D expression faces, as depicted in Figure 4. We have presented an innovative approach that combines emotionally aggressive modeling with advanced 3D face reconstruction techniques to create a specialized and unique 3D face database specifically tailored for Vietnamese individuals, known as VN3D face. By integrating these methods, our research aims to improve the accuracy and realism of facial representation, considering the distinctive characteristics of Vietnamese subjects. The resulting database (Figure 6) has significant potential for diverse applications in fields such as computer graphics, animation, and facial analysis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

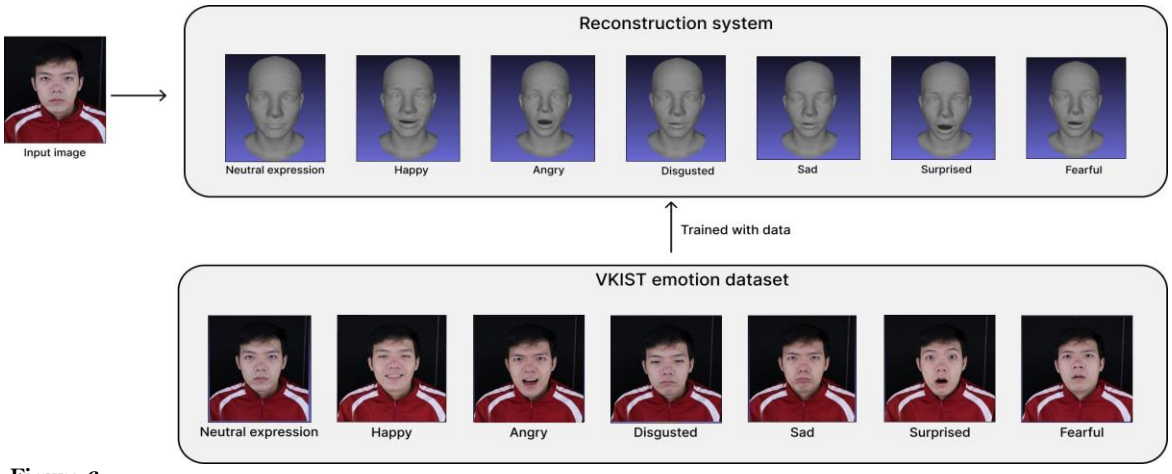

Figure 6.
Expression facial 3D face reconstruction.

The renderer takes a coarse shape (a 3D mesh) and an albedo map as input. Camera parameters $Pcam$, which correspond to capturing the frontal 3D mesh model of the face, must be provided. The output is an image, Ir, representing the shading results of the 3D mesh in a 2D format.

4. Expression Retargeting

Utilizing the 3DVNFACE database and the emotional regressive model, we have developed an emotion retargeting application, as demonstrated in Figure 7. The objective of expression retargeting involves taking two input face images: a source face image and a target face image. The desired outcome is a 3D avatar of the character portrayed in the target face image while preserving the facial expression of the character depicted in the source face image.

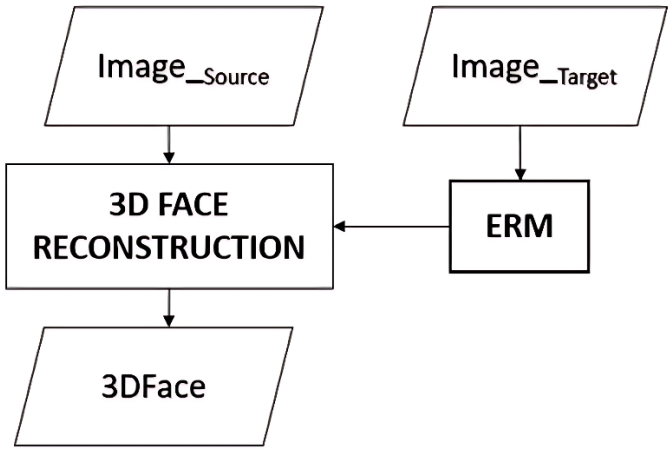

Figure 7.
Process of 3D face expression retargeting.

To achieve this, we employ the target face image to reconstruct the 3D face using DECA/MICA, as outlined in Figure 4. Subsequently, the target face image is inputted into our ERM to determine the expression and pose states using two parameters, ψ and θ. These parameters are then utilized to replace the corresponding ψ and θ parameters in the reconstruction process of the 3D face model (Figure 8).

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

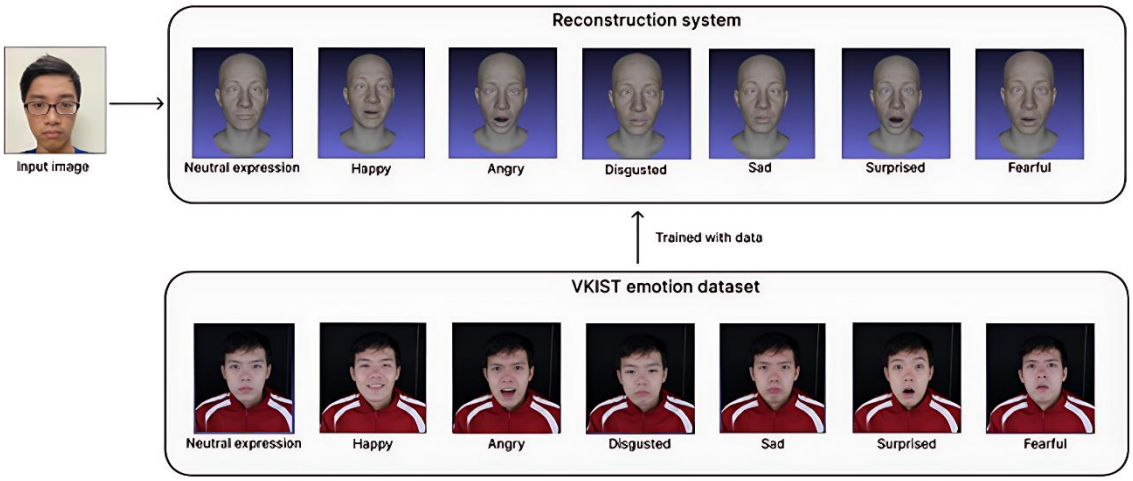

Figure 8.
3D expression retargeting.

5. Experiments

5.1. Measurement of Evaluations

PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index Measurement) are widely used non-deep learning methods that measure similarity based on specific image attributes and provide information about the similarity in terms of noise and structure. Given a noise-free m × n monochrome image I and its noisy approximation K, PSNR is defined as

$$PSNR = 10log_{10}\frac{MAX^2}{MSSE} \quad (5)$$

where, MAX is the maximum possible pixel value of the image and MSE is estimated as follow:

$$MSE = \frac{1}{mn}\sum_{i=1}^{m-1}\sum_{j=1}^{n-1}[I(i,j) - K(i,j)]^2 \quad (6)$$

Good PSNR values range from 30 to 50 dB, with higher values indicating better quality. Acceptable values fall between 20 dB and just below 30 dB.

The $\boldsymbol{SSIM}$ index is calculated on various windows of an image. The measure between two windows x and y of common size $N \times N$ is

$$SSIM(x,y) = \frac{(2\mu_x\mu_x + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (7)$$

where μ_x and μ_y are the averages of x and y, σ_x and σ_y are the variance of x and y, σ_{xy} is the covariance of x and y, c1 and c2 are two variables to stabilize the division with a weak denominator. SSIM values range from -1 to 1, with a value of 1 indicating that the two datasets are identical. A higher index value corresponds to a better model.

The **LPIPS** (Learned Perceptual Image Patch Similarity) [20]: is a deep learning-based metric that employs a neural network to learn image features and compute the similarity between two images based on these features. Two images are similar if their LPIPS score is less than 1.1.

The **DFID** (Distance in Facial Identity Feature space) [21]: is an evaluation method determining whether two portrait images belong to the same person by comparing the difference between their embedding vectors. The FaceNet framework, as defined in reference utilizes the concept of DFID to assess the preservation of identity. According to the FaceNet standard, if the DFID value between two facial images is less than or equal to 1.1, they are considered to depict the same individual.

NoWchallenge [22]: provides a benchmark method specialized in measuring the accuracy and robustness of 3D face reconstruction methods. It uses a set of 7 landmark points to rigidly align the

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

predicted mesh with the ground truth mesh. The landmark points are the leftmost and the rightmost points of the two eyes, the tip of the nose, the leftmost point, and the rightmost point of the mouth. The error is then calculated using the absolute distance between each scan vertex and the closest point in the mesh surface.

5.2. Dataset

The V-Face Dataset [18]: To train the emotion model ERM, we used the dataset from VKIST which consists of 889 images of 127 subjects with 7 captured emotions (neutral, anger, disgust, fear, happiness, sadness, surprise). All images were captured from the front view with a resolution of 2976×1984. The link to download the dataset is available. 90% of the images were used for training and 10% for testing. From the dataset, we extracted the face region and resized it to 256×256. We then used the pre-trained DECA model to extract the expression parameters from the face images. These expression parameters served as the ground truth for the emotion model.

The FaceScape dataset [10]: To evaluate the emotion model, we used the FaceScape dataset, which consists of 847 subjects. The dataset includes scanned 3D face models of 20 different expressions, and 56 images corresponding to 56 camera angles to create each of the scanned models. Only the data of 359 subjects were used because the remaining subjects were not provided with the captured images/expressions, which made it more difficult to evaluate the effectiveness of the method.

5.3. Evaluation

5.3.1. Evaluation Scenario 1

After generating the 3D expression face models within the VN3DFace dataset using images from VFace [18] (Figure 9), we proceed to project these models onto 2D images represented as Ir. Subsequently, we evaluate the quality of the 2D images using a comparative analysis against two other methods, namely i3DMM [13] and MoFaNeRF [14] employing metrics such as PSNR, SSIM, and LPIPS. Table 2 presents the results obtained from the aforementioned metrics, comparing our method with several other 3D face reconstruction techniques. The findings indicate that, in terms of direct comparison, our quantitative results are favorable compared to i3DMM and MoFaNeRF, suggesting that the VN3DFace dataset can be utilized in future studies.

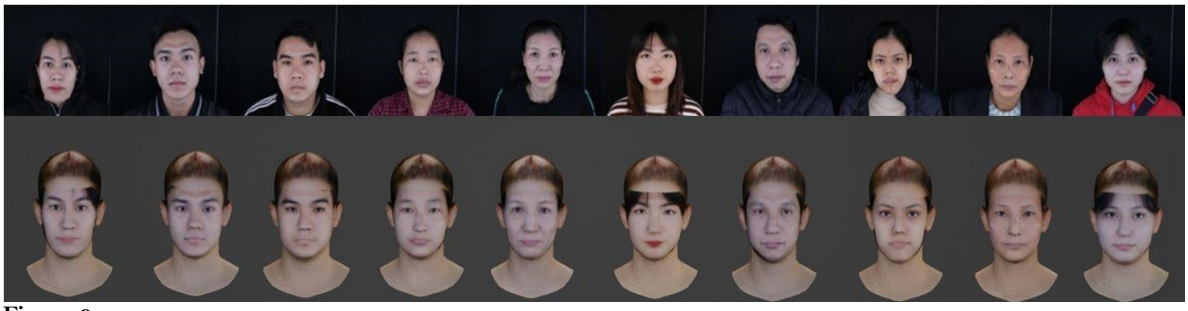

Figure 9.
V-Face 2D image input and 3D model output.

Due to outperforming certain methods in direct comparison, our approach still showcases promising results and meets the necessary quality standards. These findings highlight the potential and feasibility of our method in generating 3D face models, particularly when considering the specific context of hairless faces. Further research and refinement of our approach may contribute to achieving even more competitive results in future evaluations and comparisons.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

Table 2.
Comparison of our proposal and others.

	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	30.07	0.84	0.09
i3DMM	24.45	0.90	0.11
MoFaNeRF	31.49	0.95	0.06

5.3.2. Evaluation Scenario 2

In our evaluation, we specifically focus on assessing the representation of the same individual across the input image pairs (Ii) and the corresponding output image pairs (Ir). To measure the accuracy of this representation, we employ the DFID model, which quantifies the distance in facial identity feature space. The average DFID value obtained from our evaluation is 0.25. This value indicates that the synthesized output images have been successfully evaluated as retaining the essential features that represent the same person as depicted in the input images. The low average DFID value reinforces the effectiveness of our approach in preserving the identity and facial attributes of the individuals throughout the synthesis process. This evaluation outcome highlights the reliability and fidelity of our method in capturing and reproducing the distinctive characteristics of individuals, ensuring that their identity is accurately maintained in the synthesized output images.

5.3.3. Evaluation Scenario 3

In our evaluation, we assess the representation of the same individual exhibiting identical facial expressions across the image pairs Ii and Ir at five distinct levels: completely different, different, neutral, somewhat similar, and very similar. These levels are measured on a scale ranging from 1 to 5. We collect observational ratings from a group of individuals aged between 18 and 30 to obtain a comprehensive assessment. The survey involved 33 respondents, as depicted in Figure 10. Remarkably, the results demonstrate that an impressive 93.8% of the respondents were able to correctly identify that the input image and the synthesized face image belonged to the same person based on their observational ratings (Score ≥ 3). Out of these respondents, 14% rated the similarity between the images as" very similar" with a score of 5, while 49.3%rated the similarity as" somewhat similar" with a score of 4. These findings indicate that the synthesized face images successfully capture and retain the essential facial attributes and expressions of the individuals, allowing them to be correctly identified by most respondents. Furthermore, a substantial portion of the respondents rated the similarity between the input and synthesized images as high, indicating a close resemblance. This evaluation outcome underscores the effectiveness of our approach in accurately representing and preserving the facial features and expressions of individuals. The results validate the high-quality output of our synthesis method and highlight its potential for various applications, such as face recognition, virtual avatars, and animation.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

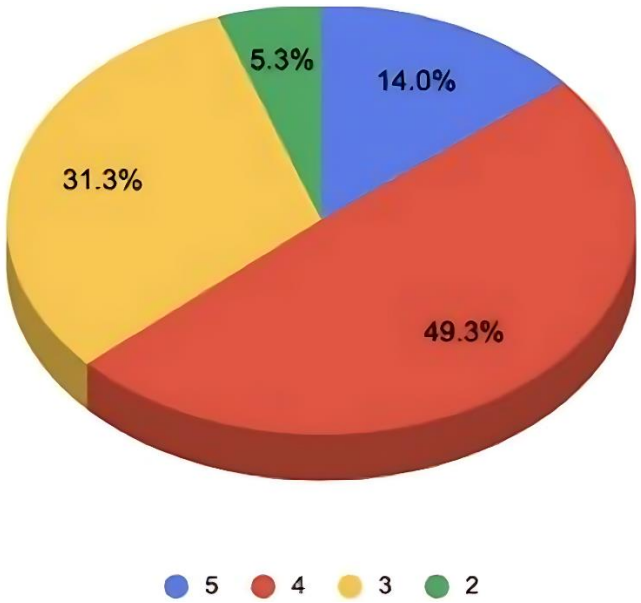

Figure 10.
Survey result.

5.3.4. Evaluation Scenario 4

In this scenario, we establish the emotional 3D faces within the FaceScape dataset as the ground truth. To generate our emotional 3D face models, we utilize both the original DECA model and our proposed method, employing the emotional 2D images available in the FaceScape dataset. Subsequently, we compare the emotional 3D face models generated by our approach with the emotional

3D face models present within the FaceScape dataset. The comparison results, as depicted in Tables 3, along with the visual representation provided in Figure 11, reveal the quality and usability of the emotional 3D face models generated by our method.

Table 3.
NOW front view and side view image input.

	Front view			Side view		
	Mean	**Median**	**Std.**	**Mean**	**Median**	**Std.**
DECA	1.83	1.46	1.49	1.77	1.40	1.44
Ours	1.75	1.45	1.41	1.78	1.42	1.46

Our evaluation findings demonstrate that the emotional 3D face models generated through our approach exhibit a high level of accuracy and fidelity when compared to the emotional 3D face models in the FaceScape dataset. This indicates that our method successfully captures and represents the emotional expressions present in the 2D images, effectively translating them into realistic and accurate 3D facial models. These results affirm the reliability and efficacy of our method, highlighting its potential for various applications that rely on emotional facial analysis and synthesis. The high-quality emotional 3D face models generated through our approach can be utilized in future endeavors, such as emotion recognition systems, virtual reality applications, and facial animation in the entertainment industry. By establishing a strong correlation between our generated emotional 3D face models and the ground truth within the FaceScape dataset, we ensure the validity and utility of our method for advancing research and practical applications in the field of emotional facial analysis and synthesis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

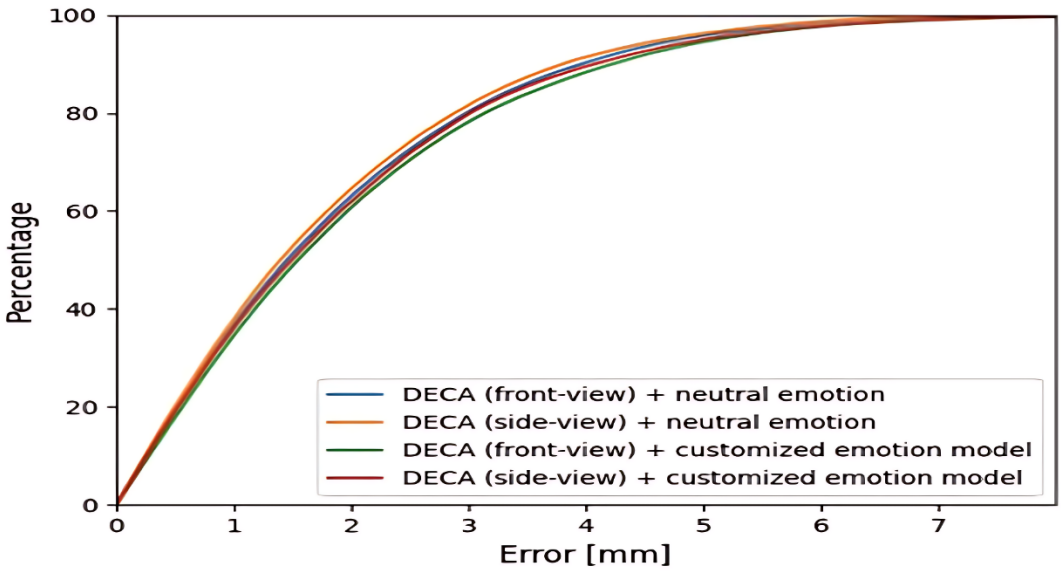

Figure 11.
Errors.

6. Conclusion

We have introduced a comprehensive framework that incorporates various methods and integrates an emotionally aggressive model to generate a high-quality 3D face dataset specifically designed for Vietnamese individuals, encompassing a wide range of facial expressions (see Figure 6). The results obtained from our proposed method have demonstrated its effectiveness in creating significantly larger and more realistic 3D datasets using only 2D face images. This breakthrough has immense implications for the field, as it opens new avenues for generating extensive and diverse 3D facial datasets without the need for expensive and time-consuming 3D scanning technologies.

With this innovative methodology, we have successfully curated a comprehensive dataset called 3DVNFace. This dataset comprises both 3D models and corresponding 2D facial images, showcasing a rich variety of expressions from 400 Vietnamese individuals. By including both types of data, we ensure a holistic representation of facial features and expressions, allowing for a more nuanced analysis and application of the dataset in various research domains. We extend an invitation to fellow researchers and professionals to utilize our dataset, which can be accessed through the dedicated website1. We believe that sharing this resource will foster collaboration, encourage further advancements in the field, and contribute to the development of more accurate and robust facial analysis algorithms and applications.

Furthermore, building upon the 3DVNFace dataset, we have developed an initial 3D retargeting application. This application offers a convenient and efficient means of transferring facial expressions between different 3D faces. By leveraging the rich data available in the dataset, this application facilitates the exploration of expression transfer techniques and starts possibilities for creating personalized avatars, enhancing virtual reality experiences, and advancing facial animation in various domains such as gaming, film, and virtual communication platforms.

In summary, our framework and the resulting 3DVNFace dataset represent a significant advancement in the field of 3D face modeling and analysis. We invite researchers to utilize this valuable resource and explore its potential applications, while also encouraging collaboration and further advancements in the field of computer vision and facial analysis.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883

Transparency:

The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Copyright:

References

[1] T. Zhang *et al.*, "Accurate 3d face reconstruction with facial component tokens, in," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[2] F. Taubner, P. Raina, M. Tuli, E. W. Teh, C. Lee, and J. Huang, "3d face tracking from 2d video through iterative dense uv to image flow, in," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[3] A. Rai *et al.*, "Towards realistic generative 3d face models," in *In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 3738-3748)*, 2024.

[4] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *arXiv:2012.04012*, 2021.

[5] W. Zielonka, T. Bolkart, and J. Thies, "Towards metrical reconstruction of human faces," presented at the In: European Conference on Computer Vision, 2022. URL https://api.semanticscholar.org/CorpusID:248177832, 2022.

[6] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero, "Learning a model of facial shape and expression from 4D scans," *ACM Trans. Graph*, vol. 36, no. 6, pp. 194:1-194:17, 2017.

[7] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques - SIGGRAPH '99*, vol. 40, pp. 187–194, 1999. https://doi.org/10.1145/311535.311556

[8] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413-425, 2013. https://doi.org/10.1109/TVCG.2013.249

[9] E. Wood *et al.*, "3d face reconstruction with dense landmarks. In European conference on computer vision." Cham: Springer Nature Switzerland, 2022, pp. 160-177.

[10] H. Yang *et al.*, "Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction," in *In Proceedings of the Ieee/cvf Conference on Computer Vision and Pattern Recognition (pp. 601-610)*, 2020.

[11] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set," *arXiv:1903.08527*, 2020.

[12] W. Paier, A. Hilsmann, and P. Eisert, "Interactive facial animation with deep neural networks," *IET Comput. Vis*, vol. 14, pp. 359–369, 2020. https://api.semanticscholar.org/CorpusID:224806560

[13] T. Yenamandra *et al.*, "i3dmm: Deep implicit 3d morphable model of human heads," in *Proceedings of the IEEE / CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[14] Y. Zhuang, H. Zhu, X. Sun, and X. Cao, "Mofanerf: Morphable facial neural radiance field," presented at the European Conference on Computer Vision, 2022.

[15] R. Danecek, M. J. Black, and T. Bolkart, "EMOCA: Emotion driven monocular face capture and animation," presented at the Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 20311–20322, 2022.

[16] B. Chaudhuri, N. Vesdapunt, L. Shapiro, and B. Wang, "Personalized face modeling for improved face reconstruction and motion retargeting," *arXiv:2007.06759*, 2020.

[17] L. Moser, C. Chien, M. Williams, J. Serra, D. Hendler, and D. Roble, "Semi-supervised video-driven facial animation transfer for production," *ACM Transactions on Graphics*, vol. 40, no. 6, pp. 1-18, 2021. https://doi.org/10.1145/3478513.3480515

[18] H. Duong, T. Tran, V.-B. Nguyen, and V. A. Dao, "V-FACE: A large-scale vietnamese face image database in unconstrained environments," *International Journal of iRobotics*, vol. 5, no. 2, pp. ˇ36-41, 2022.

[19] P. Ekman, *Basic emotions*. Sussex, UK: John Wiley Sons, 1999.

[20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," presented at the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595. https://doi.org/10.1109/CVPR.2018.00068, 2018.

[21] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] S. Sanyal, T. Bolkart, H. Feng, and M. Black, "Learning to regress 3d face shape and expression from an image without 3d supervision," in *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 1: 131-145, 2025
DOI: 10.55214/25768484.v9i1.3883