Reflections on Diversity: A Real-time Virtual Mirror for Inclusive 3D Face Transformations

Paraskevi Valergaki
Computer Science Department,
University of Crete
Institute of Computer Science,
Foundation for Research & Technology - Hellas (FORTH)
Heraklion, Greece
vivibalergaki@gmail.com

Antonis Argyros
Computer Science Department,
University of Crete
Institute of Computer Science,
Foundation for Research & Technology - Hellas (FORTH)
Heraklion, Greece
argyros@ics.forth.gr

Giorgos Giannakakis*
Institute of Computer Science,
Foundation for Research & Technology - Hellas (FORTH)
Heraklion, Greece
Department of Electronic Engineering,
Hellenic Mediterranean University
Chania, Greece
ggian@ics.forth.gr

Anastasios Roussos*
Institute of Computer Science,
Foundation for Research & Technology - Hellas (FORTH)
Heraklion, Greece
troussos@ics.forth.gr

***Abstract*—Real-time 3D face manipulation has significant applications in virtual reality, social media and human-computer interaction. This paper introduces a novel system, which we call Mirror of Diversity (MOD), that combines Generative Adversarial Networks (GANs) for texture manipulation and 3D Morphable Models (3DMMs) for facial geometry to achieve realistic face transformations that reflect various demographic characteristics, emphasizing the beauty of diversity and the universality of human features. As participants sit in front of a computer monitor with a camera positioned above, their facial characteristics are captured in real time. Our system provides a dynamic, responsive "mirror" effect, allowing the digital 3D model to follow the participant's motions, offering an immersive virtual reflection. Participants can further alter their digital face reconstruction with transformations reflecting different demographic characteristics—such as gender and ethnicity (e.g., a person from Africa, Asia, Europe). Another feature of our system, which we call "Collective Face", generates an averaged face representation from multiple participants' facial data. As each new face is processed, identity coefficients and textures are averaged to continuously update this collective face. A comprehensive evaluation protocol is implemented to assess the realism and demographic accuracy of the transformations. Qualitative feedback is gathered through participant questionnaires, which include comparisons of MOD transformations with similar filters on platforms like Snapchat and TikTok, focusing on realism, feature preservation, and faithfulness to demographic representation. Additionally, quantitative analysis is conducted using a pretrained Convolutional Neural Network that predicts gender and ethnicity, to validate the accuracy of demographic transformations. Project webpage: https://vivianval.github.io/ReflectionsOnDiversity**

***Index Terms*—3D Morphable Models (3DMMs), Generative Adversarial Networks (GANs), virtual mirror, interactive installation**

* Joint last authorship.

I. Introduction

The intersection of artificial intelligence, computer graphics, and human-computer interaction has witnessed rapid advancements over recent years, driven largely by the development of deep learning architectures like Generative Adversarial Networks (GANs) and 3D Morphable Models (3DMMs). Within this context, the ability to manipulate 3D face representations in real time is particularly compelling. Traditional 3D face modeling techniques, while accurate, often lack the flexibility to produce the full range of realistic textures that users demand. Similarly, GAN-based face generators, though capable of producing highly realistic 2D images, are not inherently designed for 3D face manipulation and struggle to maintain identity consistency. What is more, existing face manipulation applications [1]–[4], rarely address transformations based on ethnicity. Additionally, to the best of our knowledge, no face manipulation application provides a tool for users to explore realistic, ethnicity-based transformations in a meaningful way. Our approach aims to address the aforementioned limitations along with the core motivation to develop technology that fosters inclusiveness and diversity.

In this paper, we present the Mirror of Diversity (MOD) computer vision software application, as part of the *STEAM-DIVE* project, which allows users to experience realistic transformations of their faces as if they were of a different gender or ethnicity. This educational tool is in the form of a virtual interactive mirror that raises awareness around identity, multiculturalism, promoting inclusion and diversity in an educational environment.

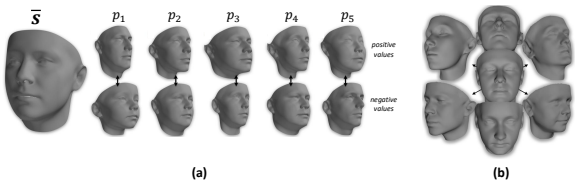

Figure 1. Visualization of the linear shape identity model of the global LSFM model: **(a)** Mean shape ($\bar{\mathbf{s}}$) and first 5 basis shapes (out of a total of n_{id}=158 basis shapes), each visualized as additions and subtractions away from the mean shape. In more detail, the top (bottom) row corresponds to setting the weight p_i of the i-th basis shape to a positive (negative) value, corresponding to 3 standard deviations of its statistical distribution. **(b)** Generation of synthetic shapes through random sampling of the shape coefficients **p**, assuming a Gaussian distribution. Figure adapted from [9].

II. Related Work

The concept of virtual mirror has been utilized in various behavioural [5], psychological [6] and healthcare applications [7], [8].

Large Scale Facial Models: Booth et al. [9] introduced the *Large Scale Facial Model (LSFM)*, a 3D Morphable Model (3DMM) trained on 12,000 high-resolution 3D scans (see Fig. 1). Compared to previous models, such as the Basel Face Model [10], LSFM exhibits greater facial variation and less tight constraints.

As 3DMMs, LSFMs consist of three components: shape identity, blendshapes, and color model. The *shape model* assumes a fixed topology T, modeling vertex coordinates as $s_{\text{id}}(p) = \bar{s} + Up$, where $\bar{s}$ is the mean shape, and U contains identity basis shapes, learned via PCA. The *blendshapes model* extends identity to capture expressions as $s(p, q) = \bar{s} + Up + Vq$, where V encodes expression variations. Similarly, the *color model* follows $c(\lambda) = \bar{c} + W\lambda$, where W contains color basis vectors.

To construct LSFM, 3D data was captured using a 3dMD™ stereo photometric device. Face meshes were aligned to establish dense correspondence, ensuring that each vertex encodes the same semantic point across samples. While *triangle meshes* are the standard representation, alternatives include *cylindrical* [11], *orthographic* [12], and *per-vertex surface normals* [13].

PCA was applied, retaining 158 principal components, explaining 99.7% of variance. LSFM captures identity variations across multiple demographics, enhanced by the *FaceWarehouse dataset* [14], which provides expression variations from 150 individuals. As high-dimensional facial shape and texture variations naturally align with demographic features, custom models were derived from MeIn3D and trained on specific demographic subsets using available data (48% male, 52% female; 82% White, 9% Asian, 5% mixed heritage, 3% Black, 1% other).

3DMM Fitting Approaches: 3DMM fitting methods primarily follow two approaches: *analysis-by-synthesis* and *deep learning*. In analysis-by-synthesis, the model parameters are adapted in an iterative manner until the synthesized shape matches the input data. Blanz and Vetter [15] pioneered an analysis-by-synthesis approach using iterative optimization techniques (e.g., gradient descent, Gauss-Newton). The goal is to estimate an optimal parameter vector P^* minimizing $E(P)$, which integrates image error, shape, texture, and rendering priors. Follow-up works framed the problem as a more general non-linear optimization, see e.g. [16]–[21].

Deep learning (DL) methods can be supervised or unsupervised. Supervised approaches train networks (e.g., regressors or encoders) to predict 3DMM coefficients from images [22]–[26]. Unsupervised methods, such as [27]–[31], are more flexible since they eliminate the need for 3D ground truth annotations.

GANs in Face Generation: Early GANs, such as Goodfellow et al. [32], generated low-resolution grayscale images. By 2018, advances like Deep Convolutional GANs (i.e. DCGANs) [33],Wasserstein GANs [34], and StyleGAN [35] enabled the synthesis of high-quality photorealistic faces. Progressive GANs (ProGANs) [36] introduced hierarchical training, stabilizing GAN convergence by growing both generator and discriminator simultaneously from low to high resolution. Their CelebA-HQ dataset significantly improved photorealism and serves as a reference in this study.

FacialGAN [37] performs style transfer while preserving identity, utilizing semantic segmentation masks to control attributes like eyes, lips, and skin. It supports interactive editing [38], allowing real-time adjustments. The framework consists of Generative, Style, Segmentation, and Discriminative Networks.

III. Methods

Figure 2 presents an overview of the proposed pipeline. It consists of seven modules. Once the calibration frames are captured, face landmarking extracts the landmarks and segmentation masks. The latter are input to FacialGAN which will produce the texture corresponding to the user's options in terms of transformations. 3DMM modules are responsible for fitting, pose and shape estimation or alteration of identity shape and therefore they are calculated at each frame. At this stage, identity coefficients along with the frontal calibration frame are sent to Collective Face. In the following sections, we provide more details about the main modules of our pipeline.

A. 3DMM Fitting

Our application employs 3DMM fitting for face reconstruction by aligning pre-constructed models with 2D facial landmarks and optionally using depth maps and texture data to generate a 3D face that closely matches the input.

3DMMs are modeled as a linear combination of identity and expression variations [39], [40]. We use LSFM models (global or bespoke), which incorporate demographic metadata [9]. Fitting involves estimating face pose and shape using an orthographic camera model, aligning extracted landmarks with their 3DMM correspondences, and solving a linear system via Singular Value Decomposition (SVD). More specifically, the 3D shape is projected into 2D by applying 3D rotation and

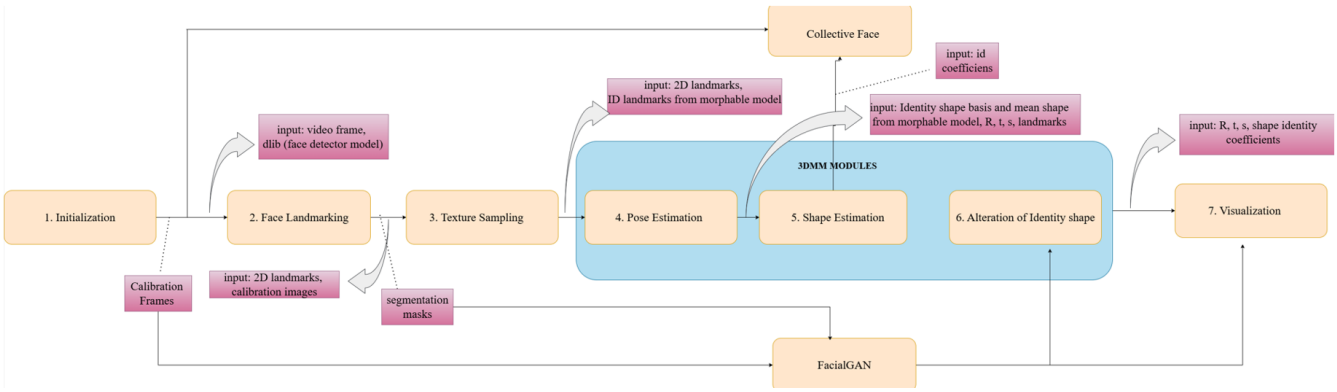

Figure 2. Pipeline of the proposed system.

translation matrices, which account for the camera's position, and then the vertices are mapped to 2D based on factors like focal length and perspective. Identity and expression coefficients are estimated by minimizing a linear least squares problem, similar to [39]. To mitigate jitter, a temporal smoothing step averages poses across three consecutive frames.

Pose Estimation: Pose estimation determines head position via a rotation matrix R, translation vector t, and scale s. Following [41], the method constructs a linear system where matrix A encodes the 3D model points, and vector d contains 2D landmarks:

$$A\mathbf{k} = d, \quad \text{where} \tag{1}$$

$$A_{2i-1} = \begin{bmatrix} u_i & v_i & w_i & 1 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{2}$$

$$A_{2i} = \begin{bmatrix} 0 & 0 & 0 & 0 & u_i & v_i & w_i & 1 \end{bmatrix}. \tag{3}$$

Solving for $\mathbf{k}$ yields the translation components $t_x = k_3/s$, $t_y = k_7/s$ and scale $s = (\|\mathbf{R}_1\|_2 + \|\mathbf{R}_2\|_2)/2$. The 3D rotation matrix is refined using Singular Value Decomposition (SVD). For temporal stability, smoothing is applied to pose estimation. The rotation matrix is updated using a weighted combination of the current and the previous frame, ensuring smoother transitions.

Shape Estimation: The goal is to estimate face shape coefficients by fitting a 3D morphable model to detected 2D landmarks. Given pose estimation parameters (R, s, t), mean shape $\bar{f}$, and principal components P, the process involves (1) constructing cumulative matrices $A = \text{proj} \cdot P$, $h = \text{landmarks} - s \cdot t - \text{proj} \cdot \bar{f}$, with $\text{proj} = s \cdot (I \otimes R_{[1:2]})$; (2) applying hyper-box constraints with $C = [I \ -I]$ and $d = \pm k \cdot \lambda$ so that we come up with plausible faces, (3) minimizing reprojection error $\|A \cdot x - h\|^2$ subject to $C \cdot x \leq d$, and (4) extracting identity coefficients α and reconstructing shape $\hat{f} = P \cdot \alpha + \bar{f}$. It should be mentioned that, for the sake of robustness, our algorithm considers only identity shape parameters during the calibration phase.

Alteration of Identity shape: The goal is to transform a participant's 3D facial geometry to resemble a target demographic.

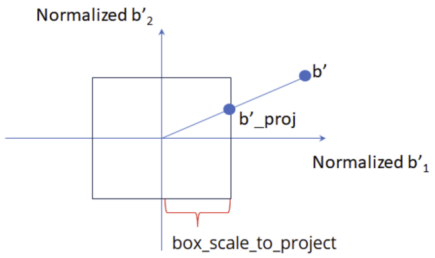

Figure 3. PCA projection of identity shape into the allowable range defined by box_scale_to_project

Given identity shape s, the bespoke model's mean shape μ', principal components U', and standard deviations σ, the pipeline involves: (1) centering the shape as $s_{\text{centered}} = s - \mu'$; (2) projecting onto PCA space as $b' = s_{\text{centered}}^T U'$; (3) applying constraints via $b'_{\text{proj}} = b' \cdot \frac{\text{box\_scale\_to\_project}}{\max(|b'|/\sigma)}$ if $\max(|b'|/\sigma) >$ box_scale_to_project; (4) reconstructing the altered shape as $s_{\text{PCAproj}} = U' b'_{\text{proj}} + \mu'$ (see Figure 3).

B. Morphing

Morphing interpolates between two 3D face states over time, transitioning from the original to a PCA-constrained version. The interpolation factor $p_{\text{morph}}(t)$ follows a sinusoidal function

$$p_{\text{morph}}(t) = \frac{1 + \sin\left(2\pi \frac{t}{T_{\text{morphing}}}\right)}{2},$$

where T_{morphing} is the total morphing period. This ensures smooth oscillation between 0 and 1, controlling the blend between the original and projected shapes. Morphing updates occur periodically based on a frame counter mechanism, ensuring seamless transitions.

C. FacialGAN

To modify facial textures, we build upon FacialGAN [37]. We adopt the MULTI-PIE markup scheme [42] with 68 facial landmarks (see Fig. 4(a)).

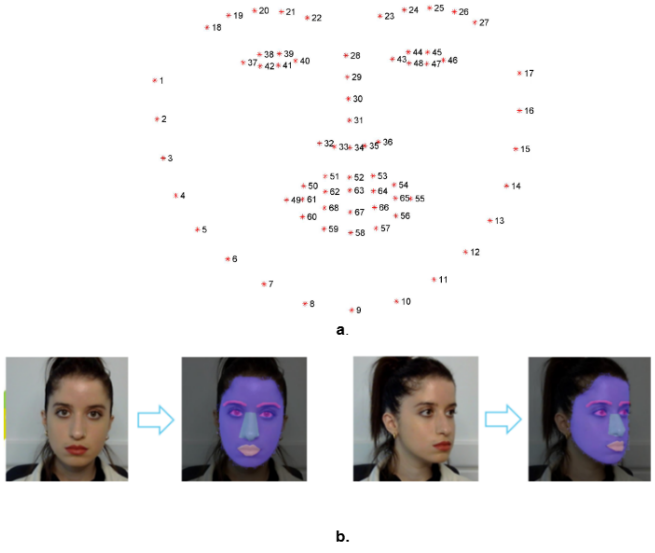

Figure 4. (a) Adopted mark-up scheme of 68 facial landmarks (Figure from [39]), (b) The process of mask generation. The first image shows the input face, the second illustrates the segmented face with different regions (mouth, nose, eyes, and eyebrows) covered by distinct masks and visualized by different colors.

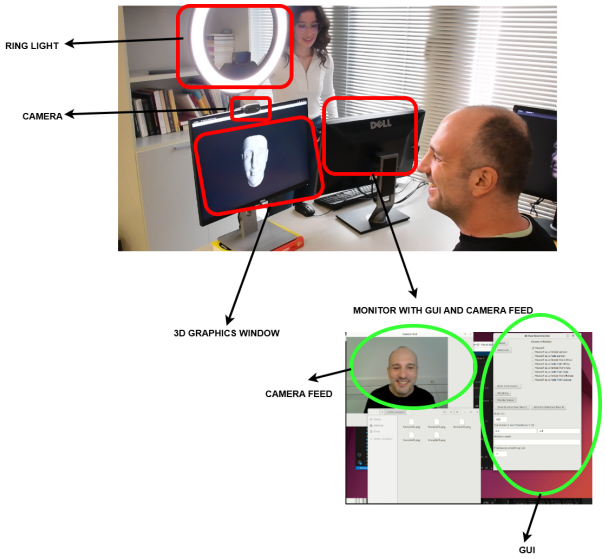

Figure 5. Installation setup with GUI, camera, lighting, and OpenGL rendering.

The alignment process begins by computing the eye centers, where the left eye center is $\mathbf{C}_{\text{left}} = \frac{1}{6}\sum_{i=37}^{42}(x_i, y_i)$ and the right eye center is $\mathbf{C}_{\text{right}} = \frac{1}{6}\sum_{i=43}^{48}(x_i, y_i)$. The rotation angle is then calculated as $\theta = \arctan\left(\frac{y_{\text{right}} - y_{\text{left}}}{x_{\text{right}} - x_{\text{left}}}\right)$, while the scale factor is determined as $s = \frac{d_{\text{desired}}}{\sqrt{(x_{\text{right}} - x_{\text{left}})^2 + (y_{\text{right}} - y_{\text{left}})^2}}$, ensuring proper spatial consistency.

Using these parameters, we calculate a rigid transformation M to bring all faces in a common reference space:

$$\mathbf{M} = \begin{bmatrix} s\cos\theta & -s\sin\theta & t_x \\ s\sin\theta & s\cos\theta & t_y \end{bmatrix}$$

and apply it to the original image I, yielding the aligned face I_{aligned}. Similarly, each landmark $\mathbf{p}_i = (x_i, y_i)$ is transformed using $\mathbf{p}'_i = \mathbf{M} \cdot \mathbf{p}_i$.

To generate segmentation masks for distinct facial regions (e.g., nose, mouth, eyes), we use the transformed facial landmarks as well as a UNet-based segmentation [43] (see Fig. 4(b)). The final aligned and segmented face and masks are sent to FacialGAN, producing high-quality 256×256 transformed images. Reference images are selected from CelebA-HQ [44] for each bespoke transformation. According to the modification type determined by the user, an interpolation function for smooth texture blending that combines manipulated and non-manipulated textures is applied.

D. Collective Face

The *Collective Face* feature generates an averaged face from multiple participants' data. Facial landmarks are extracted, aligned, and identity coefficients are averaged for continuous updates. The collective face is periodically sent to FacialGAN via Dropbox for texture or expression manipulation. A Graphical User Interface (GUI) enables users to assign their transformation to Collective Face F (female) or Collective Face M (male), with future support planned for non-binary options.

E. Implementation

The installation includes a computer, two monitors (one for the instructor, one for the participant), a Logitech C270 HD webcam, and a ring light for uniform illumination. The system processes the live feed and mirrors it for the virtual reflection effect. The GUI allows real-time calibration, adjusting scaling and translation parameters for precise alignment (Figure 5).

Our system was tested on an AMD Ryzen 9 7900 (12-core, 24-thread) processor, 61 GiB RAM, and NVIDIA RTX 3060 for GPU-accelerated processing.

IV. Experiments

The evaluation of our Mirror Of Diversity system's face reconstruction and ethnicity transformations is conducted through a combination of quantitative and qualitative methods, aiming to assess both the technical accuracy and the user-perceived realism and accuracy of the transformations.

A. Qualitative Analysis

To provide an initial overview of the capabilities of the MOD software, this section presents illustrative examples of its functionalities. Figure 6 showcases the capabilities of the MOD software in performing facial transformations across different genders and ethnicities while providing both realistic and structural visualizations. Each row represents a transformation applied to the input face, including gender changes (e.g., male to female) and ethnicity-based modifications (e.g., female from Africa, male from Asia). For each transformation, three outputs are provided: (1) input face, (2) 3D shape with texture, and (3) 3D shape only. The MOD software also supports transformations to European male and female.

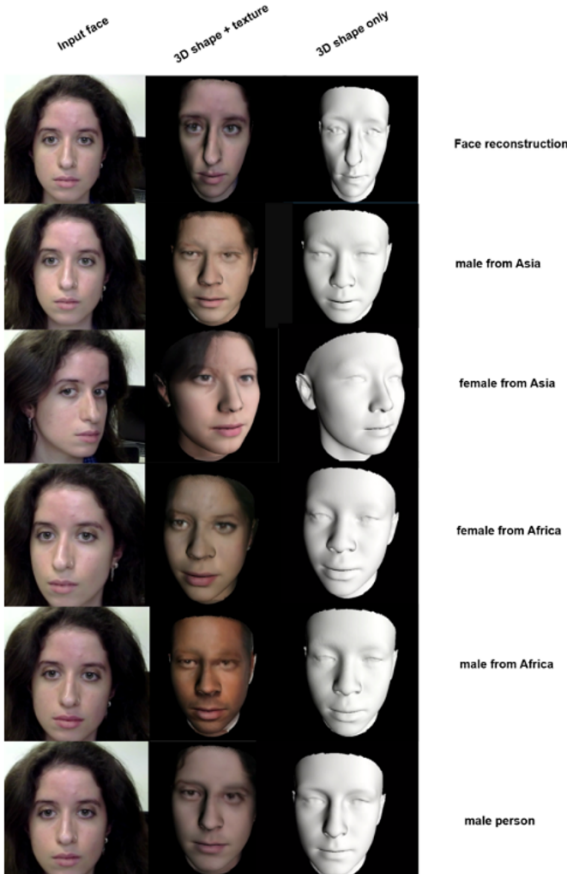

Figure 6. Examples of input and transformed faces using the MOD software. The transformations include face reconstruction, gender-based transformations (male and female), and ethnicity-based transformations (Asian and African). Each transformation is visualized in three formats: the input face, 3D shape with texture, and 3D shape only. A demo video showcasing all the functionalities of our system is available at https://vivianval.github.io/ReflectionsOnDiversity.

B. Quantitative Evaluation

We evaluate MOD's transformations using the FairFace model [45], which employs a pretrained ResNet-34 CNN for gender and ethnicity classification. We compare MOD-generated transformations to face-manipulation filters from Snapchat and TikTok, as well as a version of the Real-Time 4D Face Reconstruction (RT4Dface system) [46], which was demonstrated during the European Researcher's Night at FORTH, Heraklion, Crete (2019). Figure 7 showcases manipulations from our system and compared methods.

To assess transformation accuracy, we use FairFace's classifier to predict gender and ethnicity from OpenGL-rendered MOD images and their counterparts from Snapchat and TikTok. The idea behind this evaluation is that, if a method's face transformation is successful, then an automatic classifier like FairFace should classify the face as belonging to the desired class (e.g. male from Asia). The classifier outputs confidence scores for four ethnicity classes (White, African, Asian, Indian) and gender scores. The dataset includes original and transformed images (gender-swapped, male from Africa, female from Africa, male from Asia, female from Asia). Tables I - II illustrate the acquired metrics. The *macro average* is computed as Macro Avg $= \frac{1}{C}\sum_{i=1}^{C}\text{Metric}_i$, treating all classes equally, while the *weighted average* considers class distribution as Weighted Avg $= \frac{\sum_{i=1}^{C}(\text{Support}_i \times \text{Metric}_i)}{\sum_{i=1}^{C}\text{Support}_i}$, where Support_i is the number of samples in class i, emphasizing classes with more samples.

Table I
GENDER CLASSIFICATION SCORES

Metric	MOD (Ours)	Mobile Filters
F1-Score (Female)	**0.91**	0.89
F1-Score (Male)	**0.92**	0.67
Accuracy	**0.92**	0.83
Macro Avg (F1)	**0.92**	0.78
Weighted Avg (F1)	**0.92**	0.85

Table II
ETHNICITY CLASSIFICATION SCORES

Metric	MOD (Ours)	Mobile Filters
F1-Score (White)	0.62	**0.80**
F1-Score (African)	**0.67**	0.00
F1-Score (Asian)	0.40	**0.40**
Accuracy	**0.58**	0.50
Macro Avg (F1)	**0.56**	0.40
Weighted Avg (F1)	**0.56**	0.40

MOD generates transformed faces that yield higher gender classification accuracy (92%) compared to mobile filters (83%), with significantly better male transformation accuracy (F1-Score: 0.92 vs. 0.67). For ethnicity classification, MOD surpasses mobile filters in detecting African and Asian features, though there is area for improvement in these categories.

C. User Study

We conducted an anonymous online survey with 20 participants to evaluate realism, demographic accuracy, and feature preservation in video transformations. Each participant reviewed 32 video pairs, comparing MOD's transformations with those from Snapchat, TikTok, and RT4Dface [46]. Transformations were randomized to ensure unbiased evaluations.

Participants rated each transformation on a *5-point Likert scale* based on: 1. **Demographic Accuracy**: Resemblance to the intended demographic. 2. **Realism**: How realistic the transformation appears. 3. **Feature Preservation**: How much the transformation resembles the original person. Weighted averages were calculated to provide a quantitative comparison of user perceptions across the different methods.

Analysis of Bespoke Models: Among demographic transformations, male from Africa scored highest in resemblance (3.4), while male from Asia had the lowest user ratings (2.4). Female from Africa achieved the highest realism (3.05), while male from Asia was the least convincing (2.375). Feature preservation remained moderate, with male/female transformations leading at 2.475 (see Figure 8).

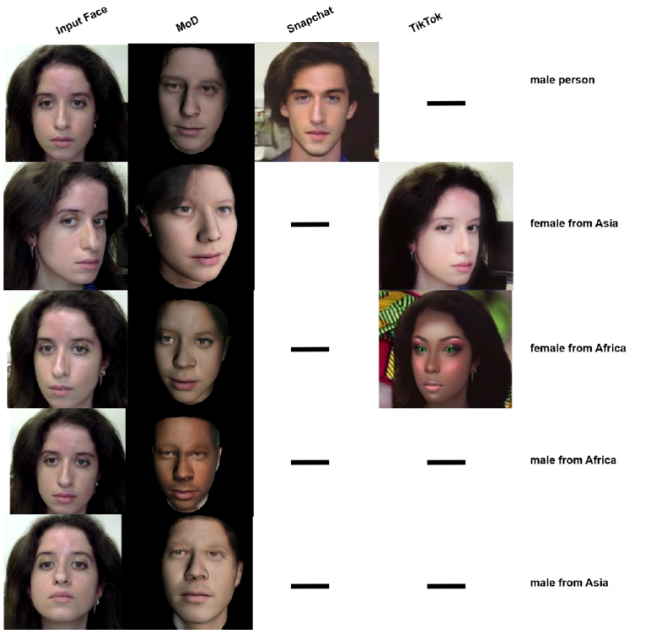

Figure 7. Comparison of how MoD, Snapchat and TikTok handle specific demographic and gender transformations. In each row the first image corresponds to the input original face and then each column is associated with the method used for the transformation. First row presents transformations to a male person achieved with MoD and Snapchat.

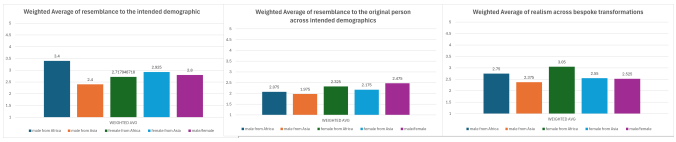

Figure 8. Female and male from Africa have in general higher user ratings in terms of resemblance to intended demographic and realism

Comparative Analysis: MOD significantly outperformed RT4Dface, achieving a weighted average of 3.54 for demographic resemblance (vs. 1.74). The most improved transformation was male from Asia, increasing from 1.5 to 3.55. Realism also improved, scoring 3.27 vs. 1.46 in RT4Dface with a notable shift from lower-rated categories (”Not at all” or ”Slightly Accurate”) to higher-rated ones (”Moderately Accurate” or above) as depicted in Figure 9.

Compared to Snapchat and TikTok, MOD’s transformations had lower demographic resemblance (2.8 vs. 3.2) and realism (2.6 vs. 2.9), likely due to real-time expression effects in filters. However, MOD performed better in female from Asia (demographic resemblance) and female from Africa (realism), highlighting its strength in culturally specific transformations (see Figure 10).

V. Conclusion and Future Research

This study demonstrates MOD’s effectiveness in face reconstruction and ethnicity-based transformations, outperforming RT4Dface in realism (3.275 vs. 1.4625) and demographic

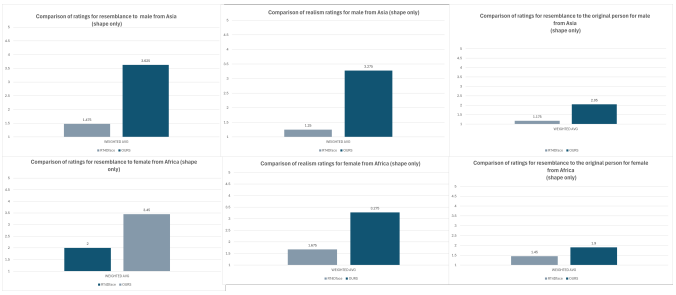

Figure 9. MOD outperforms RT4Dface in realism, demographic resemblance, and identity preservation.

Figure 10. MOD surpasses mobile filters in specific cases as in realism for female from Africa and resemblance to female from Asia.

resemblance (3.537 vs. 1.7375). The system excels in African transformations but struggles with male from Asia, indicating areas for refinement.

While Snapchat/TikTok generally scored higher, MOD outperformed in specific cases, such as female from Asia (demographic resemblance) and female from Africa (realism). These findings highlight MOD’s strength in fine-grained demographic modeling. Future improvements should focus on (1) real-time facial dynamics (blinking, directional gaze) for enhanced realism, (2) Improved lighting robustness to reduce inconsistencies, (3) expanding beyond binary gender transformations for greater inclusivity, (4) refining demographic models to represent more specific cultural identities, (5) enhancing recall for African and Asian transformations via larger datasets and (6) extending demographic transformations to finer-grained cultural representations (e.g., countries instead of continents). These refinements will ensure MOD produces more accurate, diverse, and representative transformations across populations.

Acknowledgments

This work was co-funded by the project STEAMDIVE: Diversity in STEAM, Innovative Educational Tools for Promoting Inclusion and Diversity in Schools, supported by the European Commission under the ERASMUS+ Programme, KA220-SCH - Cooperation partnerships in school education (GA: 2022-1-EL01-KA220-SCH-000086968).

This work was also co-funded by the Hellenic Foundation for Research and Innovation (HFRI) under the "1st Call for HFRI Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment", project I.C.Humans, no 91.

REFERENCES

[1] B. News. (2016) Snapchat under fire for 'yellowface' photo filter. Accessed: 2024-11-19. [Online]. Available: https://www.bbc.com/news/world-asia-37042475

[2] B. Newsbeat. (2016) Snapchat bob marley filter accused of promoting blackface. Accessed: 2024-11-19. [Online]. Available: https://www.bbc.com/news/newsbeat-36310925

[3] ——. (2016) Snapchat criticised for 'beautifying' coachella filter. Accessed: 2024-11-19. [Online]. Available: https://www.bbc.com/news/newsbeat-36091249

[4] J. Bryner, "Neanderthal app turns you into a caveman," *NBC NEWS*, May 2010. [Online]. Available: https://www.nbcnews.com/id/wbna37095461

[5] C. M. Grewe, T. Liu, A. Hildebrandt, and S. Zachow, "The open virtual mirror framework for enfacement illusions: Enhancing the sense of agency with avatars that imitate facial expressions," *Behavior research methods*, vol. 55, no. 2, pp. 867–882, 2023.

[6] Y. Inoue and M. Kitazaki, "Virtual mirror and beyond: The psychological basis for avatar embodiment via a mirror," *Journal of Robotics and Mechatronics*, vol. 33, no. 5, pp. 1004–1012, 2021.

[7] Y. Andreu-Cabedo, P. Castellano, S. Colantonio, G. Coppini, R. Favilla, D. Germanese, G. Giannakakis, D. Giorgi, M. Larsson, P. Marraccini *et al.*, "Mirror mirror on the wall… an intelligent multisensory mirror for well-being self-assessment," in *2015 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2015, pp. 1–6.

[8] Y. Andreu, F. Chiarugi, S. Colantonio, G. Giannakakis, D. Giorgi, P. Henriquez, E. Kazantzaki, D. Manousos, K. Marias, B. J. Matuszewski *et al.*, "Wize mirror-a smart, multisensory cardio-metabolic risk monitoring system," *Computer Vision and Image Understanding*, vol. 148, pp. 3–22, 2016.

[9] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large Scale 3D Morphable Models," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 233–254, 2018.

[10] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009, pp. 296–301.

[11] J. J. Atick, P. A. Griffin, and A. N. Redlich, "Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images," *Neural Computation*, vol. 8, no. 6, pp. 1321–1340, 1996.

[12] R. Dovgard and R. Basri, "Statistical symmetric shape from shading for 3d structure recovery of faces," in *Computer Vision - ECCV 2004*, T. Pajdla and J. Matas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 99–113.

[13] O. Aldrian and W. A. P. Smith, "Inverse rendering of faces on a cloudy day," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 201–214.

[14] 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, 2014.

[15] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99. USA: ACM Press/Addison-Wesley Publishing Co., 1999, p. 187–194. [Online]. Available: https://doi.org/10.1145/311535.311556

[16] H. Li, J. Yu, Y. Ye, and C. Bregler, "Realtime facial animation with on-the-fly correctives." *ACM Transactions on Graphics*, vol. 32, no. 4, pp. 42–1, 2013.

[17] C. Cao, Q. Hou, and K. Zhou, "Displaced dynamic expression regression for real-time facial tracking and animation," *ACM Transactions on Graphics*, vol. 33, no. 4, pp. 1–10, 2014.

[18] P. Garrido, M. Zollhöfer, D. Casas, L. Valgaerts, K. Varanasi, P. Pérez, and C. Theobalt, "Reconstruction of personalized 3d face rigs from monocular video," *ACM Transactions on Graphics*, vol. 35, no. 3, pp. 1–15, 2016.

[19] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Transactions on Graphics*, vol. 34, no. 6, 2015.

[20] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," *Communications of the ACM*, vol. 62, no. 1, p. 96–104, 2018.

[21] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Facevr: Real-time facial reenactment and eye gaze control in virtual reality," *arXiv preprint arXiv:1610.03151*, 2016.

[22] K. Olszewski, J. J. Lim, S. Saito, and H. Li, "High-fidelity facial and speech animation for vr hmds," *ACM Transactions on Graphics*, vol. 35, no. 6, 2016.

[23] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen, "Production-level facial performance capture using deep convolutional neural networks," 2017. [Online]. Available: https://arxiv.org/abs/1609.06536

[24] H. Kim, M. Zollöfer, A. Tewari, J. Thies, C. Richardt, and C. Theobalt, "Inversefacenet: Deep single-shot inverse face rendering from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[25] M. R. Koujan, A. Roussos, and S. Zafeiriou, "Deepfaceflow: In-the-wild dense 3d facial motion estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[26] M. C. Doukas, M. R. Koujan, V. Sharmanska, A. Roussos, and S. Zafeiriou, "Head2head++: Deep facial attributes re-targeting," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 3, no. 1, p. 31–43, Jan. 2021. [Online]. Available: http://dx.doi.org/10.1109/TBIOM.2021.3049576

[27] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3d morphable model regression," *CoRR*, vol. abs/1806.06098, 2018. [Online]. Available: http://arxiv.org/abs/1806.06098

[28] A. Tewari, M. Zollöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and C. Theobalt, "MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction," in *Proceedings of the International Conference on Computer Vision*, 2017.

[29] 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," in *Proceedings of the Workshops of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[30] L. Tran and X. Liu, "On learning 3d face morphable model from in-the-wild images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 157–171, 2019.

[31] M. B. R, A. Tewari, H.-P. Seidel, M. Elgharib, and C. Theobalt, "Learning complete 3d morphable face models from images and videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

[32] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014. [Online]. Available: https://arxiv.org/abs/1406.2661

[33] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2016. [Online]. Available: https://arxiv.org/abs/1511.06434

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein gan," 2017. [Online]. Available: https://arxiv.org/abs/1701.07875

[35] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," 2019. [Online]. Available: https://arxiv.org/abs/1812.04948

[36] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," 2018. [Online]. Available: https://arxiv.org/abs/1710.10196

[37] R. Durall, J. Jam, D. Strassel, M. H. Yap, and J. Keuper, "Facialgan: Style transfer and attribute manipulation on synthetic faces," 2021. [Online]. Available: https://arxiv.org/abs/2110.09425

[38] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," 2020. [Online]. Available: https://arxiv.org/abs/1907.11922

[39] J. Deng, A. Roussos, G. Chrysos, E. Ververas, I. Kotsia, J. Shen, and S. Zafeiriou, "The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking," *Int. J. Comput. Vision*,

vol. 127, no. 6–7, p. 599–624, Jun. 2019. [Online]. Available: https://doi.org/10.1007/s11263-018-1134-y

[40] J. Booth, A. Roussos, E. Ververas, E. Antonakos, S. Ploumpis, Y. Panagakis, and S. Zafeiriou, "3d reconstruction of "in-the-wild" faces in images and videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 11, p. 2638–2652, Nov. 2018. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2832138

[41] A. Bas, W. Smith, T. Bolkart, and S. Wuhrer, "Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences," 02 2016.

[42] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vision Comput.*, vol. 28, no. 5, p. 807–813, May 2010. [Online]. Available: https://doi.org/10.1016/j.imavis.2009.08.002

[43] F. P. Papantoniou, P. P. Filntisis, P. Maragos, and A. Roussos, "Neural emotion director: Speech-preserving semantic control of facial expressions in "in-the-wild" videos," 2022. [Online]. Available: https://arxiv.org/abs/2112.00585

[44] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proceedings of the International Conference on Learning Representations*, 2018.

[45] K. Kärkkäinen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age," 2019. [Online]. Available: https://arxiv.org/abs/1908.04913

[46] M. R. Koujan, N. Dochev, and A. Roussos, "Real-time monocular 4d face reconstruction using the lsfm models," 2020. [Online]. Available: https://arxiv.org/abs/2006.10499