IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 21 July 2025, accepted 29 July 2025, date of publication 1 August 2025, date of current version 8 August 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3594875

RESEARCH ARTICLE

SynAdult: Multimodal Synthetic Adult Dataset Generation via Diffusion Models and Neuromorphic Event Simulation for Critical Biometric Applications

MUHAMMAD ALI FAROOQ, (Senior Member, IEEE), PAUL KIELTY, WANG YAO, AND PETER CORCORAN, (Fellow, IEEE)
School of Engineering, University of Galway, Galway, H91TK33 Ireland

Corresponding author: Muhammad Ali Farooq (muhammadali.farooq@universityofgalway.ie)

This work was supported in part by the Taighde Eireann–Research Ireland under Grant IRCLA/2023/1992; and in part by the ADAPT–Centre for Digital Content Technology, Enterprise Ireland.

ABSTRACT We propose SynAdult, a multimodal synthetic data generation framework designed to address the scarcity of diverse and privacy-compliant senior adult face datasets for biometric applications and facial analysis. Our pipeline begins with the rendering of high-fidelity 2D adult facial images using parameter-efficient LoRA-based tuning of the state-of-the-art hyperrealism Stable Diffusion XL (SDXL) model, producing photorealistic outputs across diverse ethnicities and age-specific features. Next, we integrate a video retargeting pipeline to synthesize temporally consistent head pose and facial expression sequences, ensuring naturalistic dynamics crucial for downstream video-based facial analysis. In the third stage, we generate neuromorphic event data to introduce a privacy-preserving modality aligned with real-world edge deployment scenarios, such as ambient monitoring and in-vehicle sensing, where high temporal resolution and minimal identity leakage are beneficial. Finally, we reconstruct detailed 3D facial meshes from single 2D frames using 2D-to-3D morphing techniques to capture fine-grained structural details. This modality enhances geometric understanding and supports applications in AR/VR and affective computing. To validate the robustness and utility of the generated dataset, we perform a comprehensive evaluation using Kernel Inception Distance (KID), BRISQUE, CLIP score, and identity similarity metrics. We further assess downstream applicability employing the state-of-the-art facial expression classification networks and event facial landmarks tests for downstream machine learning tasks. As a key contribution, we open-source a large-scale, multimodality, multi-race adult dataset, enabling future research in secure and ethically grounded synthetic data for facial biometrics and facial analysis applications. The project website, along with the complete adult multimodality dataset and the fine-tuned model, is available at https://mali-farooq.github.io/SynAdult/

INDEX TERMS Multimodality, synthetic data, diffusion models, neuromorphic event imaging, privacy, V2E, ML.

I. INTRODUCTION

With the accelerating pace of aging, there is a growing demand for medical care, health monitoring, and emotional support tailored to older adults. Facial expression recognition (FER) systems offer a promising approach for assessing emotional states, facilitating the early detection of potential mental health concerns among the elderly [1]. Prior studies [2] have demonstrated that facial expressions can effectively be used to monitor and document patients' pain

The associate editor coordinating the review of this manuscript and approving it for publication was Deepak Mishra.

levels. Furthermore, research [3] validated the applicability of facial expression analysis in mental health, providing a non-invasive way for monitoring emotional well-being.

However, age-related changes in facial anatomy, such as increased wrinkling, skin laxity, and reduced muscle tone, can significantly hinder the accurate interpretation of facial expressions. Sönmez et al. [4] examined the impact of age-specific training data on the performance of facial expression recognition models. Their findings indicated that recognizing expressions in older adults posed the greatest difficulty among all age groups. Ko et al. [5] found that there are differences in facial expression intensity and muscle usage across age groups. For instance, older adults were more prone to expressing negative emotions and exhibited increased muscle engagement in the lower jaw region compared to younger individuals. Thus, addressing the challenges associated with facial expression analysis in aging populations, especially in studies focusing on expression recognition between ages, is an important research question.

A key contributing factor is the limited availability of high-quality facial expression datasets specifically featuring older adults. As highlighted by [6], the current datasets lack adequate representation of this demographic, which significantly impedes the progress of facial expression recognition technologies. Moreover, even in controlled experimental environments, capturing consistent expressions across diverse individuals remains inherently difficult due to natural variations in expressive behavior.

Recent advances in facial analysis have been significantly driven by the availability of high-quality synthetic datasets, which have enabled robust training of deep learning models across a range of tasks such as recognition, verification, and expression classification. Notably, works like [7] and [8] have demonstrated how controlled data generation can alleviate dataset bias, expand identity coverage, and boost downstream performance. However, despite this progress, facial analysis in older adults remains under explored and more challenging due to the increased intra-class variation introduced by aging such as skin texture changes, wrinkles, and structural sagging. These age-induced variances degrade the performance of conventional models trained predominantly on younger faces, as previously shown in studies [9] focusing on age interval-based benchmarks. Further complexity arises in the context of facial expression analysis. Unlike identity recognition, expression generation and classification suffer from low inter-subject consistency. That is, one individual's smile or frown may appear visually very different from another's, even under controlled conditions. Additionally, demographic representation issues arise, as datasets frequently underrepresent certain racial and gender groups, leading to biased models that perform poorly on underrepresented populations. Last but not least ethical concerns and GDPR further complicate data collection, including obtaining informed consent, protecting privacy, and ensuring cultural sensitivity during the process. These challenges collectively limit the diversity, quality, and fairness of datasets, resulting in ML models that fail to generalize effectively, particularly for adult expression analysis in real-world applications.

Keeping these challenges in view, we have proposed a multimodal senior adult dataset created using latent diffusion models by employing text-to-image and image-to-video generation methods. This approach enables the synthesis of a baseline dataset that captures distinct and granular facial expressions across multiple races and genders. This work is part of our broader research on domain-specific face data synthesis for underrepresented demographics. While this study focuses on senior adults, we have also developed synthetic datasets for children and facial aging, demonstrating their utility for robust, face data analysis and bias-aware biometric systems [7], [8], [9], [10]. Using the generative capabilities of diffusion models, we address the issues of underrepresentation and demographic bias inherent in traditional data collection methods. This synthetic dataset incorporates nuanced variations in facial expressions and aging-related characteristics, ensuring wider inclusivity and diversity. The proposed solution aims to provide a robust foundation for downstream machine learning tasks, particularly in adult expression analysis, by offering a high-quality, balanced, and ethically sourced dataset. To perform this task, we utilized a hyperrealism cinematic-style XL model as the base architecture and further fine-tuned it on real-world adult datasets. This fine-tuning process allowed the model to accurately capture the intricate details and variations found in senior adult faces, such as wrinkles, skin textures, and age-specific expressions. By combining the hyperrealistic generative capabilities of the base model with domain-specific training, we generated robust and high-quality 2D images of senior adults. This approach ensures that the synthetic data maintains a high degree of realism while addressing the diversity required for effective downstream machine learning applications.

To address both data scarcity and ethical concerns around privacy in facial data collection, we also explore the integration of neuromorphic vision as part of our synthetic multimodal pipeline. Event-based cameras, which asynchronously capture changes in pixel intensity rather than full RGB frames, offer a alternative for privacy-aware sensing. These optical sensors produce sparse, low-latency data streams that inherently obscure high-resolution personal identity traits, making them attractive for applications in driver monitoring systems and in-home healthcare, especially for older adults who may require non-intrusive observation. Motivated by these emerging use cases, we incorporated the V2E (Video-to-Event) simulator into our framework to synthesize realistic event data aligned with RGB and 3D modalities. This addition supports the development and evaluation of learning algorithms that can operate under privacy-constrained conditions while maintaining performance in dynamic environments. This step enhances the dataset by adding temporal dynamics, which are crucial for

tasks involving real-time tracking and motion detection. The V2E simulator enabled the conversion of synthetic video animations into event-based data, capturing pixel-level changes over time with high temporal resolution. By integrating this neuromorphic data stream, our approach not only enriches the dataset's diversity but also provides a comprehensive resource for training and evaluating multimodal machine learning models in various real-world scenarios.

Lastly, we focused on generating 3D data from the synthesized 2D images, enabling the creation of detailed and realistic 3D models that capture facial geometry and spatial structure. Using 2D-to-3D reconstruction pipelines, 3D data retain the diversity and realism of the synthetic data while offering a robust resource for tasks requiring 3D representations, such as facial expression analysis, avatar creation, and biometric modeling. By incorporating 3D files into the dataset, we ensure its utility for applications in VR/AR environments, 3D animation, and other scenarios that require high-quality 3D data.

A. MAIN CONTRIBUTIONS OF THIS PAPER

The core contributions of this research are as follows.

- We proposed and generated a diverse, inclusive dataset using hyperrealism cinematic-style SDXL models, fine-tuned on real-world adult datasets. This dataset incorporates multiple races (Asian, African and White), genders (male and female) and a wide range of facial expressions animations to address the limitations of traditional datasets, such as demographic underrepresentation and bias.
- By using the state-of-the-art V2E event simulator, we added a neuromorphic event data stream to capture dynamic temporal changes, complementing the visible video data. Additionally, we extend the dataset by generating 3D face structures from a single 2D generated image, thus providing detailed spatial representations.
- We performed extensive validation tasks to ensure the dataset's quality and reliability, including facial landmark validation and large language model (LLM)-based facial expression analysis. This comprehensive validation ensures that the synthetic data accurately reflects real-world scenarios.
- To further drive innovation, we have open-sourced the SynAdult dataset and tuned diffusion model weights,[1] allowing researchers and developers to build and evaluate machine learning models with diverse, ethically sourced data.

II. BACKGROUND/RELATED WORK

The recent advancements in Generative AI particularity diffusion models have increased innovation in the field of computer imaging and high-end visual synthesis, enabling the creation of hyperrealistic images, videos, and 3D models with high fidelity. Stable Diffusion is a powerful latent diffusion model capable of creating detailed and diverse outputs from textual prompts, offering fine-grained control over generated content. Its strengths lie in its ability to generate realistic images with fine textures, intricate details, and diverse stylistic choices, while maintaining efficiency and scalability. The model operates in a latent space, which significantly reduces computational costs compared to traditional pixel-based approaches, enabling faster and more scalable image inference process. This flexibility makes Stable Diffusion ideal for inpainting, data augmentation, and simulation by generating realistic, diversified controlled, and reproducible synthetic data to build more targeted AI models.

A. FACE DATA SYNTHESIS USING TEXT-TO-IMAGE DIFFUSION

This section will highlight recent studies exploring the capabilities of Stable Diffusion, in generating photo realistic human face data across various applications. Liao et al. [11] addressed the challenge of rendering high-quality, realistic human faces by using text-to-image generation approach. Authors fine-tuned diffusion models to improve facial detail rendering, introducing a novel metric called Face Score (FS) to better align with human judgments of face quality. Their approach demonstrated significant improvements in generating photorealistic faces from textual descriptions. Ergasti et al. [12] proposed a Semantic Latent Diffusion Model for human face generation and editing. Their framework integrates shape and style information through SPADE normalization and cross-attention layers, allowing for precise control over each semantic part of the human face. This method enables both the reproduction and manipulation of real reference images, as well as the generation of diverse high-quality facial images. Banerjee et al. [13] explored the use of latent text-to-image diffusion models for synthetically aging and de-aging face images. They trained there models, using few-shot learning, and maintained high degrees of visual realism while preserving biometric identity. Their approach has shown potential applications in improving automated face recognition systems by providing high-quality aged face datasets. Rowan et al. [14] addressed the challenge of accurate 3D face reconstruction by generating a large-scale synthesized dataset named SynthFace. By conditioning Stable Diffusion on depth maps samples from the FLAME 3D Morphable Model, they produced a diverse set of shape-consistent facial images. Their model, ControlFace, trained on SynthFace, achieved competitive performance without requiring 3D supervision or manual 3D annotations. For realistic avatar creation, authors in [15] proposed Human 3Diffusion, a framework that combines 2D multi-view diffusion models with 3D reconstruction techniques. By integrating these models, they achieved realistic avatars from single RGB image, ensuring both geometric and appearance accuracy. This method underscores the potential of diffusion models in generating precise 3D human representations.

[1] Website: https://mali-farooq.github.io/SynAdult/

B. MULTIMODALITY DATA SYNTHESIS

Several studies have explored the generation of multimodal faces from visual attributes, with recent advancements in deep learning techniques to enhance image synthesis quality. Authors in [16] Att2MFace, a novel network architecture designed to generate multimodal face images from visual attributes using a unified generator. Unlike existing approaches that rely on multiple models or complex pipelines, Att2MFace employs a single generator capable of producing diverse facial images while maintaining consistency across different modalities. Additionally, the model benefits from a progressive training strategy, which enhances the synthesis of coherent and high-fidelity multimodal face images. To achieve effective multimodal fusion, authors in [17] propose a lightweight fusion architecture that exploits variance across channels at different feature layers. This approach seamlessly integrates with existing models, enabling zero-shot generation, which would otherwise be infeasible with a single model. Additionally, authors demonstrate that their method extends beyond spatial conditioning and is equally effective for style-based conditioning. In [18] authors proposed a comprehensive framework for 3D facial animation synthesis from speech, addressing challenges posed by the scarcity of high-quality 4D facial data and multimodal annotations. They introduce Generalized Neural Parametric Facial Asset (GNPFA), a variational autoencoder that maps facial geometry and images into a generalized expression latent space, effectively decoupling expressions from identity. Using GNPFA, the authors construct M2F-D, a large-scale scan-level 3D facial animation dataset with rich emotion and style labels by extracting high-quality expressions and head poses from diverse video sources.

C. SYNTHESIZING NEUROMORPHIC EVENT DATA

Event cameras offer significant advantages for face analysis [19], [20], revolutionizing the way facial dynamics are captured and interpreted. Their high temporal resolution and low-latency operation enable real-time tracking of facial expressions, lip movements, and eye gestures, making them ideal for applications such as human-computer interaction and critical computer vision applications. Unlike traditional cameras, which may fail to capture rapid facial transitions, event cameras can detect subtle and fast-changing facial dynamics. Additionally, their wide dynamic range ensures accurate facial feature representation even in challenging lighting conditions, outperforming conventional cameras that often struggle with overexposure or underexposure. Further, a key advantage of neuromorphic vision lies in its ability to support privacy-preserving sensing. Event cameras, which capture sparse and asynchronous pixel-level changes, offer a robust solution for scenarios where conventional image-based monitoring may raise ethical or privacy concerns. In line with ongoing efforts, which emphasize the development of controllable and adaptable AI models for deployment in privacy-sensitive, edge-based environments (e.g., in-home elder monitoring or transportation systems), we incorporate event-based data generation as part of our multimodal synthesis pipeline. This inclusion not only aligns with practical deployment challenges but also supports the training of AI models that are robust to temporal dynamics while offering improved privacy guarantees.

Event-based vision simulators [21] have shown good potential and play a crucial role in generating synthetic event data, enabling the use of such data for various downstream machine learning applications without requiring physical event cameras. These simulators convert standard frame-based video into event streams by estimating changes in pixel intensity, thus replicating the output of real event cameras.

One widely used simulator is Video-to-Event (V2E), introduced by Hu et al. [22], which converts conventional videos into realistic event streams by computing per-pixel intensity changes. Another prominent tool is Event-based Synthetic Image Generator (ESIM) by Rebecq et al. [23], which offers precise event simulation with adjustable noise modeling and realistic physics-based scene synthesis. Additionally, in [24] authors introduced simulation methods that enhance event simulation performance by two orders of magnitude, achieving real-time capability while maintaining robust data quality.

In this study, we utilized V2E to convert video animations of adult aging into event data streams, leveraging its ability to simulate neuromorphic vision through asynchronous spike-based encoding. By capturing per-pixel luminance changes over time, V2E generates a temporally precise and high-dynamic-range event representation, effectively preserving fine-grained aging-related facial transformations such as skin texture evolution, and structural deformations. This enables efficient encoding of temporal aging dynamics, reducing redundant frame-based information while enhancing data fidelity for applications in age-progressive facial synthesis, biometric aging models, and age-invariant recognition systems.

D. 3D FACE MORPHING

The transformation of 2D facial images into 3D representations has gained significant attention due to its applications in computer vision, biometrics, healthcare, and digital identity protection. In particular, adult aging face synthesis plays a crucial role in age progression studies, forensic investigations, and medical applications. Recently, deep learning has significantly improved the morphing of 2D to 3D aging faces, leading to more accurate, detailed, and realistic reconstructions.

3D face reconstruction is commonly used to recreate the 3D shape, expressions, and textures of a face from 2D images. In 1999, Blanz et al. [25] introduced 3D Morphable Models (3DMM), a pioneering approach for generating facial shape and appearance. Their work laid the foundation for 3D face reconstruction by enabling the estimation of 3DMM parameters to derive a textured mesh that best

aligns with an input 2D face. The introduction of Neural Radiance Fields (NeRFs) [26] provided a novel way to represent 3D faces with photorealistic detail, capturing complex light interactions and view-dependent effects. In the realm of texture generation, Marriott et al. [27] combined Progressive GANs [28] with 3DMM to enhance facial textures for recognition; however, their approach struggled to preserve high-frequency details. A significant improvement was later achieved by Slossberg et al. [29], who leveraged StyleGAN [30] and StyleRig [31] to generate high-resolution textures with better realism and consistency. Despite these advancements, the high cost of obtaining real-world data makes large-scale data collection challenging, limiting the availability of training data and, consequently, the development of models. To address this issue, FFHQ-UV [32] leveraged HIFI3D++ [33] and StyleFlow [34] to synthesize a high-quality UV dataset. Similarly, UV-IDM [35] adopted a comparable approach, utilizing StyleGAN2 [30] to generate the BFM-UV dataset, effectively mitigating the shortage of training data.

In this work, we have chosen UV-IDM [35] to generate 3D texture for its high-fidelity texture generation powered by the Latent Diffusion Model (LDM) and its ability to preserve identity using an Identity-Conditioned Module (ICM). The ICM ensures identity consistency by leveraging any in-the-wild image as a conditioning input. UV-IDM is also easily adaptable to BFM-based methods, making it a flexible choice for 3D face reconstruction. With its state-of-the-art performance and efficient texture generation within seconds, UV-IDM is ideal for our texture-completion task.

III. PROPOSED FRAMEWORK

Our synthetic dataset generation framework is structured into four modular pipelines as shown in Figure 1, each targeting a specific modality and contributing to the creation of a comprehensive multimodal dataset:

1) **Aged Synthetic Face Generation:** We generate high-quality synthetic facial images of older adults using text-to-image diffusion models guided by descriptive prompts. This step allows the creation of diverse seed identities across varying age ranges and ethnic backgrounds.
2) **Facial Expression Animation:** To capture temporal dynamics and expressive cues, the static aged faces are transformed into short facial video sequences. This is achieved using a video retargeting pipeline that superimposes a range of realistic facial expressions onto the synthetic identities.
3) **Event Stream Synthesis:** The generated facial videos are converted into neuromorphic event-based data using the V2E simulator [22]. These event streams offer high temporal resolution and are particularly well-suited for privacy-preserving scenarios in real-time edge applications.
4) **2D-to-3D Morphing:** To enhance the structural representation of facial features, we reconstruct 3D face meshes from the original 2D synthetic images. This enables downstream tasks that require geometric reasoning or multimodal 2D–3D fusion.

These modular pipelines collectively enable the synthesis of a large-scale, demographically diverse, and privacy-aware dataset suitable for a variety of AI-driven facial analysis applications. Figure 1 shows comprehensive block diagram representation of the adapted SynAdult methodological framework to generate multi-modality adult face data which is further explained in sub sections.

A. CONDITIONED SENIOR ADULT FACE SYNTHESIS USING BASE MODEL TRAINING

The first phase of the proposed algorithm fine-tunes the Stable Diffusion XL (SDXL) model for hyper-realistic text-to-image generation using a combination of Contrastive Language-Image Pretraining (CLIP) [36], Low-Rank Adaptation (LoRA) [37], Conditional Variational Autoencoder [38], and multi-dataset regularization. This process enhances the model's ability to generate photo-realistic images by incorporating diverse large scale datasets, noise-conditioning mechanisms, and efficient hyper parameter adaptation. Further VAE encoder maps high-dimensional image data into a lower-dimensional latent representation, reducing computational complexity and enabling stable training process. During generation, the VAE decoder reconstructs the final high-resolution image from the latent space, ensuring fine-grained image details. As shown in figure 1 block 1 the diffusion model was trained on three large-scale publicly available datasets which includes FFHQ, UTK and Age DB face datasets, enabling it to learn a wide range of facial features with gender, age and race attributes along with a diverse set of facial expressions. During the training phase, we have used tailored text prompts to further guide (condition) the model's attention, ensuring the generation of highly detailed and contextually accurate facial gender and race ethnicity. This combination of large-scale face datasets and curated text prompts helps in an improved generation of face data of different genders, races, and facial expressions that reflect both global context and fine-grained details.

In our specific context, targeted generation of adult synthetic faces under multimodal constraints (e.g., specific demographics, expressions, and paired text/image conditions), the pretrained SDXL model lacks the fine-grained control required to consistently meet these semantic and structural objectives. To address this, we incorporate Low-Rank Adaptation (LoRA) modules and domain-specific fine-tuning for the following key reasons:

1) **Domain Adaptation:** SDXL is trained on a broad distribution of natural images and general prompts. To ensure the model adapts to adult face data with specific biometric and text-conditional cues (e.g., age, expression, pose), we fine-tune it on curated data samples acquired from three distinct large-scale public datasets (FFHQ, UTK, and AgeDB) along with tailored prompts aligned with our use case.

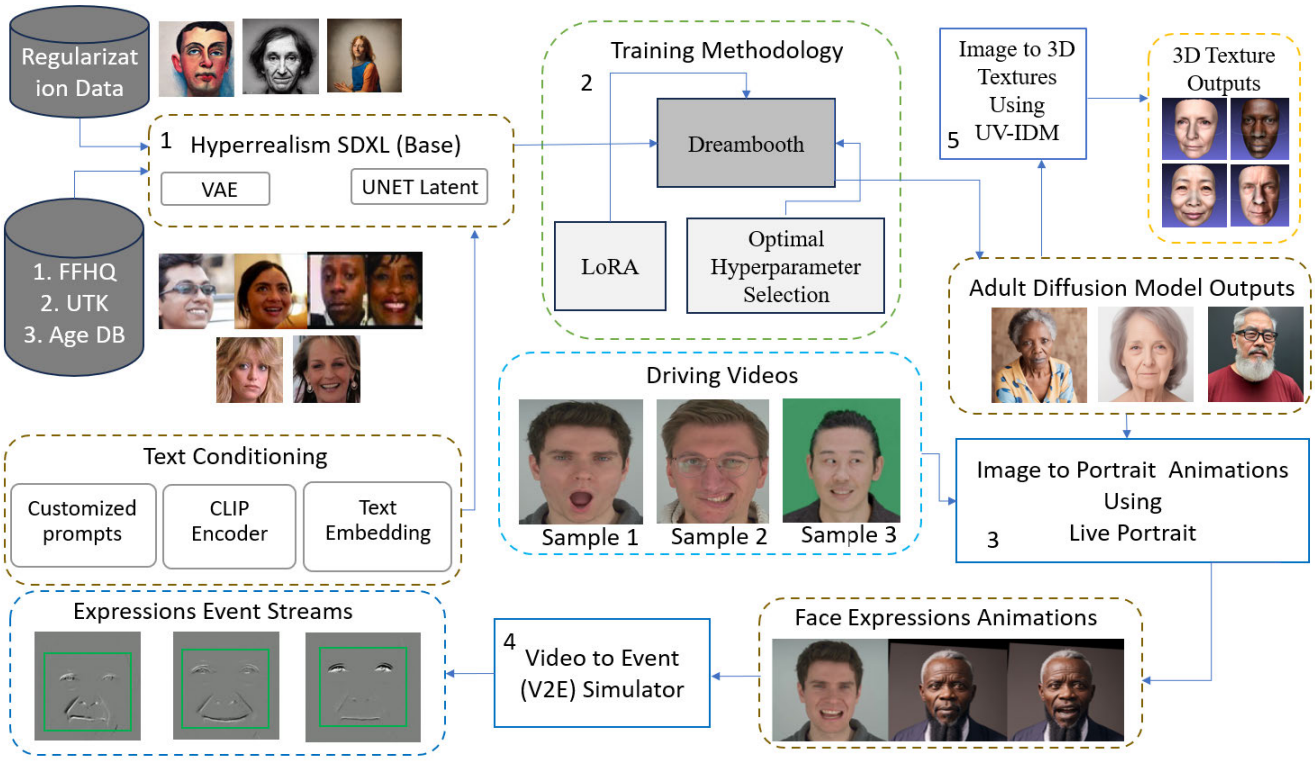


FIGURE 1. Comprehensive block diagram representation for building high-level SynAdult framework.

2) **Multimodal Alignment:** LoRA enables efficient adaptation without overfitting the base SDXL model. It allows us to inject custom embeddings and modality-specific signals (including caption-conditioned generation, attribute-driven prompts) in a parameter-efficient manner, improving controllability and image-text alignment.
3) **Visual Fidelity under Constraints:** While SDXL excels at general photorealism, its performance degrades when asked to synthesize consistent, high-fidelity facial attributes across multiple modalities (e.g., senior adult gender diversity data, diversified facial expressions, and paired annotations). Fine-tuning with LoRA improves stability and consistency in these scenarios.
4) **Parameter Efficiency and Speed:** Finally, LoRA provides a computationally efficient alternative to full fine-tuning of diffusion models. By training only a small number of additional low-rank matrices, LoRA significantly reduces memory and compute requirements, enabling fast iteration without compromising performance, which is especially valuable for large diffusion models.

In addition to this, we have also integrated a subset of person-class regularization data into the training pipeline as shown in figure 1. This dataset serves a crucial role in the regularization process, enhancing the stability and performance of Stable Diffusion XL models. By introducing greater data diversity and mitigating biases, it help to improve generalization during training. This ensures that the model does not overfit to specific individuals or characteristics, leading to more robust and balanced image generation.

In this work, we have used DreamBooth tool [39] for fine-tuning our model, enabling personalized adaptation while preserving the integrity of the pretrained Stable Diffusion XL framework. Additionally, we integrated a Low-Rank Adaptation (LoRA) model as shown in figure 1 block 2 for low-rank decomposition into weight updates. This helps to minimizes the combined loss and optimize the fine-tuning process, significantly reducing computational resource and overall training time requirements.

$$\Delta W = AB \tag{1}$$

where W is the original model weight matrix, and A, B are low-rank matrices reducing parameter updates.

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{VAE}} + \lambda_2 \mathcal{L}_{\text{denoise}} + \lambda_3 \|\Delta W\|^2 \tag{2}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters balancing each term. The model parameters θ are updated using the AdamW optimizer:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda \theta_t \right) \tag{3}$$

where:

- η is the learning rate.
- m_t, v_t are first and second moment estimates.
- λ is the weight decay parameter.

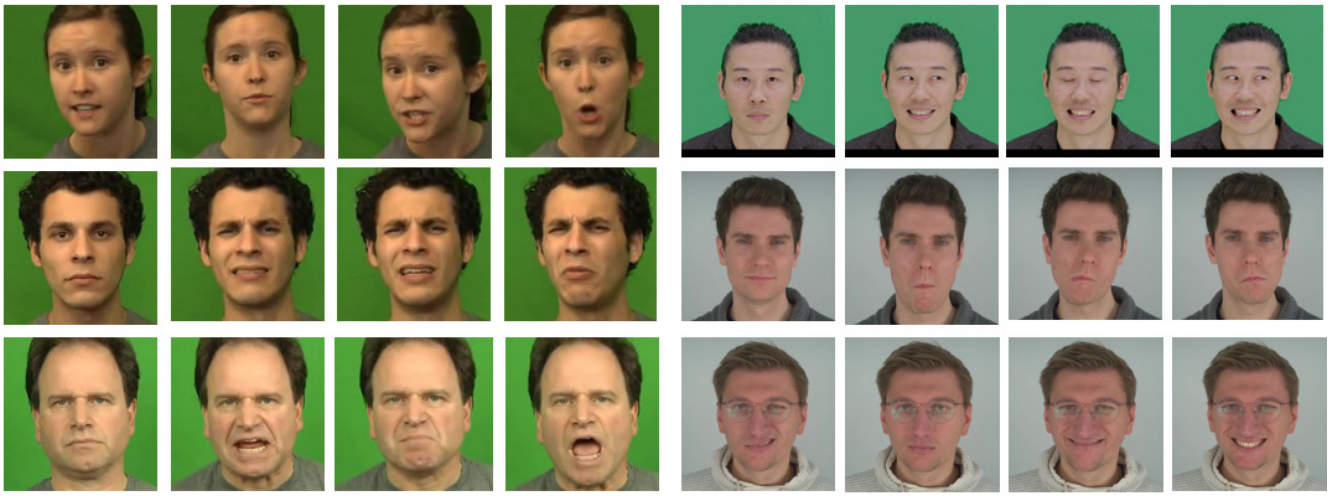


FIGURE 2. Driving video samples with distinct facial expressions and head pose variations extracted from CREMA-D and LivePortrait datasets.

Algorithm 1 shows the complete description for building Adult-Diffusion text to image model. To improve the robustness of the model, we have used Discrete Denoising Scheduler (DDS) that introduces controlled Gaussian noise into the training images. The U-Net architecture, which serves as the backbone of SDXL's denoising model, predicts noise $\hat{\epsilon} = G(I_{\text{noisy}}, C, \theta)$ and removes the noise step-by-step. U-Net plays a crucial role in refining the image generation process by leveraging skip connections, allowing for efficient multi-scale feature extraction.

Algorithm 1 Hyper Realism SDXL Fine-Tuning Algorithm With CLIP, LoRA, and Multi-Dataset Regularization

Require: Text descriptions T, Image datasets $\mathcal{D} = \{\text{FFHQ, AgeDB, UTKFace}\}$, Regularization dataset D_r.
Preprocess text descriptions T and image datasets $\mathcal{D}$;
Initialize SDXL model parameters (θ) with LoRA adaptation;
Training Steps:
  Encode the text descriptions T using CLIP text encoder ι
  Conditioning vector $C = \iota(P)$ generated using a text encoder ι and a text prompt P
  Encode the image data using an image encoder ε
  Combine the encoded text ι and image features
  Define a noise scheduler δ_t (Discrete Denoising Scheduler - DDS)
  Generate initial images using the SDXL generator network $G(T, \theta)$
  Add Gaussian noise: $I_{\text{noisy}} = I + \delta_t \cdot \mathcal{N}(0, 1)$
  Predict noise: $\hat{\epsilon} = G(I_{\text{noisy}}, C, \theta)$
  Compute loss: $\mathcal{L} = \|\hat{\epsilon} - \epsilon\|^2$
  Update model parameters (θ) using AdamW optimizer and backpropagation
  Repeat steps 5 to 13 for multiple iterations
Ensure: $I = G(T, \theta) + \mathcal{E}$, Fine-tuned SDXL model for text-to-image generation.

B. IMAGE TO PORTRAIT ANIMATION

The next stage after generating high-quality 2D portraits via the fine-tuned SDXL model, is integrating efficient image-to-video diffusion framework with enhance stitching and retargeting control to animate these images. For this purpose, we experimented with various SoA models and selected LivePortrait [40] as the backbone, for realistic facial motion (head pose) and expression synthesis as shown in figure 1 block3. This approach maps motion trajectories while ensuring that the original model retains its generalization capabilities for customized portrait animation tailored to specific input conditions by providing distinct set of reference videos.

We build upon the LivePortrait framework [40], which integrates a canonical implicit keypoint detector (L), head pose estimation network (H), and expression deformation estimator (Δ) into a unified model (M), using ConvNeXt-V2-Tiny [41] as the backbone.

$$M(I) = (L(I), H(I), \Delta(I)) \tag{4}$$

While the original architecture is optimized for general facial animation, we specifically modify the video retargeting pipeline to enable robust domain adaptation for senior adult face animation synthesis. This is done by introducing a pre-alignment module to better normalize facial geometry in aged faces and further fine-tuning the keypoint detection head using a curated set of synthetic senior adult faces to better localize age-specific landmarks (such as deep nasolabial folds, sagging jawlines). These enhancements enable improved motion realism, better temporal consistency, and higher fidelity in animated outputs when working with age-progressed or elderly synthetic data. Our pipeline is particularly effective when deployed on synthetic datasets generated via diffusion models, ensuring compatibility across appearance and motion domains.

In addition, to improve image generation quality, the SPADE decoder [42] is used as the generator (G). The SPADE

decoder demonstrates superior performance compared to the original face vid2vid decoder [43], effectively leveraging the warped feature volume (f_s), where each channel functions as a semantic map for guiding the synthesis of animated images.

$$I_{\text{animated}} = G(f_s) \tag{5}$$

$$f_s = W(I, L, H, \Delta) \tag{6}$$

where $W(\cdot)$ represents the warping function that aligns the input image I based on the detected keypoints, head pose, and expression deformations.

For improved efficiency, a PixelShuffle [44] layer is integrated as the final stage of G, enabling resolution upscaling from 256 × 256 to 512 × 512 while preserving overall image quality.

$$I_{\text{final}} = \text{PixelShuffle}(I_{\text{animated}}) \tag{7}$$

C. DRIVING VIDEOS FOR FACE MOTION AND EXPRESSIONS SYNTHESIS

In this study we have used different and diversified sets of driving videos for animated style transferring of head pose and expressions on rendered 2D adult data subjects. Keeping this in view we have used short video clips with various expressions and head pose variations from LivePortrait, and CREMA-D [45] dataset as shown in Figure 1 driving videos block and Figure 2. The CREMA-D dataset comprises 7,442 audiovisual clips from 91 actors (48 male and 43 female) aged 20 to 74, representing diverse racial and ethnic backgrounds, including African American, Asian, Caucasian, Hispanic, and unspecified groups ensuring variability in expression characteristics. Moreover, the LivePortrait also provides high-quality, short video clips designed for facial reenactment tasks, offering a range of facial expressions and head pose dynamics. By utilizing these datasets, we ensure a diverse and representative set of driving signals for animated style transfer, allowing for robust adaptation across different ethnic groups and genders.

D. PORTRAIT TO EVENT DATA SIMULATION

Events were simulated using V2E as shown in Figure 1 block 4 with mostly default parameters. The only changes were a reduction of the positive and negative thresholds from 0.2 to 0.15, and the frequencies of the shot noise and lowpass cutoff were both increased from 0 to 0.5 Hz and 15 Hz respectively. Lower contrast thresholds allow the capture of more subtle motions but also adds to the number of noise events. With a limited number of driving videos, the increased noise provides valuable variance to the dataset. Hence the further addition of some shot noise.

Most of the driving videos were saved at 30 FPS, but some others were found to be 25 FPS. If events were simulated from these videos without alteration, the resulting event streams would have different temporal resolutions. Certain networks could learn to use these differing resolutions as a predictive feature, but one that is not found in real events. To mitigate this, we utilized the Super-Slomo network [46] with a dynamic upsampling factor such that all videos were interpolated to 150 FPS. This ensures the final event streams have a consistent temporal resolution of 6.67ms.

E. 2D-3D FACE SYNTHESIS

The final stage of our proposed pipeline integrates 2D-to-3D face rendering, as illustrated in block 5 of Figure 1. In this work, we have followed UV-IDM [35] to generate 3D texture from our 2D images. The UV-IDM is a high-fidelity facial texture generator designed to reconstruct realistic and identity-preserving UV textures from single in-the-wild face images. The core idea is to treat texture generation as a texture completion task, where an incomplete UV texture extracted from a face image is used as a condition for generating a high-quality full-face texture.

Firstly, a Variational Autoencoder (VAE) [47] is used to compress high-dimensional UV textures into a low-dimensional latent space. This makes training more efficient while maintaining high quality. Then, the trained VAE encoder is used to map UV textures to latent space. A Latent Diffusion Model (LDM) [48] is then trained in this space to learn the distribution of high-quality UV textures. Since traditional texture generation methods often fail to preserve identity, UV-IDM introduces an Identity-Conditioned Module (ICM) to guide the diffusion process. This ensures that the diffusion model generates textures that match the person's identity while filling in missing regions. By leveraging the ICM and LDM, we effectively preserve identity and intricate details in the generated textures. Additionally, the use of a VAE for perceptual compression provides a compact yet detailed latent space representation, enabling efficient and accurate texture generation suitable for various 3D face reconstruction tasks.

IV. EXPERIMENTAL RESULTS

The complete experimental work was carried out on a workstation machine equipped with 64GB of RAM and A6000 GPU having 48 GB of dedicated video graphics memory. Pytorch library was used for the environmental setup.

A. HIGH QUALITY ETHNIC ADULT DATA CURATION

During the training phase we conducted experiments using Dreambooth, by selecting the optimal set of training parameters and three different datasets as discussed in Algorithm 1. We have selected Discrete Denoising Scheduler (DDS) as the noise scheduler. Further, we have used 8 bit, AdamW optimizer. The training data was divided into two main subclasses (class 1: male, class 2: female). Each class included approximately 2k images, making a total of around 4k images used for DreamBooth fine-tuning. All images were preprocessed to a resolution of 512 × 512 pixels, which is the native input size supported by SDXL-based diffusion pipelines. The full DreamBooth fine-tuning procedure took approximately 4 hours, using mixed-precision training and a learning rate of $2e^{-6}$ with a batch size of 2. Incorporating

FIGURE 3. Adult male data results: First row show the white race with distinct facial identities and with and without beard, second row shows african race results, and third row shows the Asian race results with and with out beard.

FIGURE 4. Adult female data results: First row show the white race with distinct facial identities and skin texture patterns, second row shows african race results, and third row shows the asian race results with and without beard.

LoRA helps reduce video memory footprints, leading to more efficient resource utilization. The model requires approximately 7.3 GB of VRAM during the training phase. The next part includes generating photo-realistic adult facial data by deploying the tuned model in inference pipeline. Figure 3 and 4 shows rendered results with different facial characteristics and ethnic identities. During the inference phase, the samples were generated using the Euler and Euler A sampling methods, with a sampling scale ranging from 20 to 28 and a CFG scale of 7. The rendered samples showcase detailed facial features, including skin texture, fine wrinkles, and subtle expressions. This level of realism makes them ideal for applications like facial recognition, analysis, and animation.

B. EXPRESSIONS AND HEAD POSE ANIMATIONS

After rendering 2D images for both male and female subjects, we used a distinct set of driving videos in LivePortrait to generate facial animations. Given a driving video sequence $\{D_i \mid i = 0, \ldots, N-1\}$, we applied style transfer to our 2D static images by extracting motion features from each

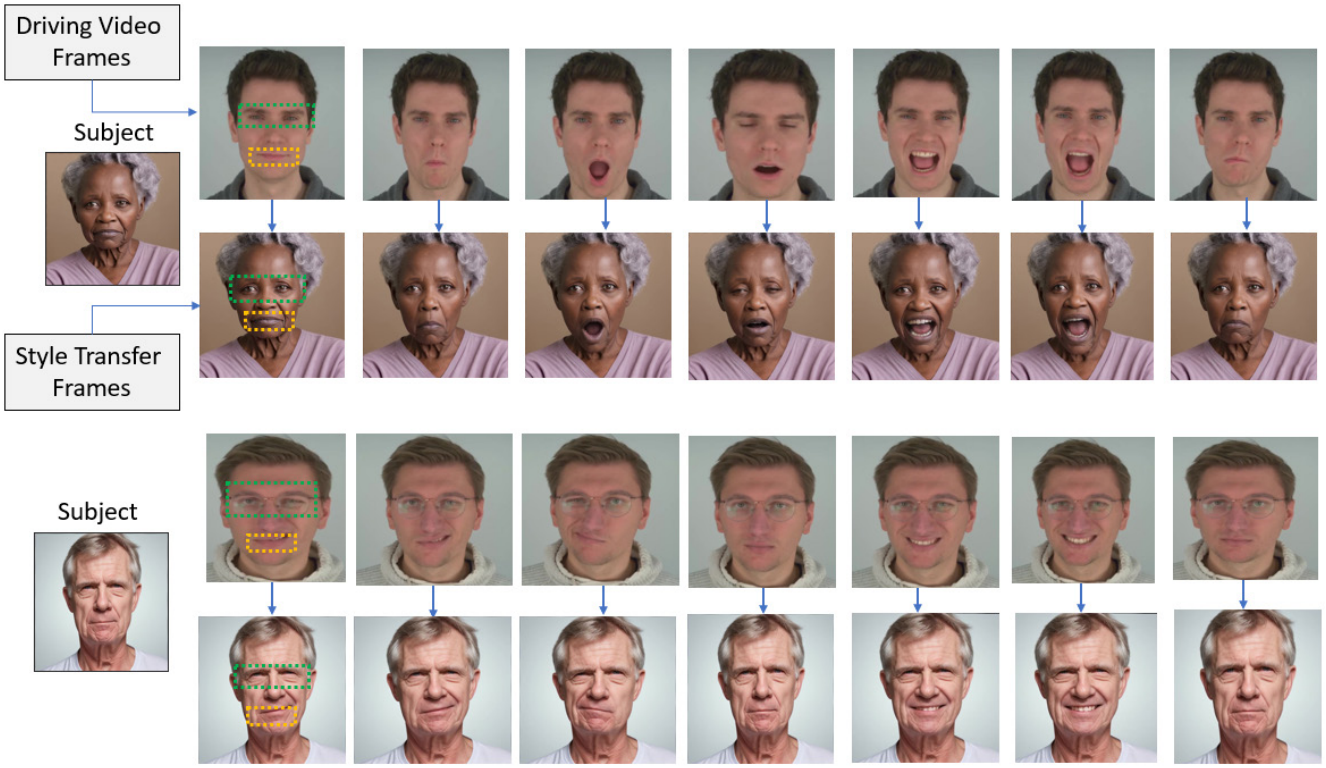


FIGURE 5. Style transfer results on two different subjects, demonstrating a range of expressions, micro-expressions, facial muscle movements, as well as dynamic eye and lip motions.

frame, including scale factor S_i, deformation parameters D_i, translation vector T_i, and rotation matrix $R_i = M(D_i)$. The scale factor S_i ensures proportionality in the facial region, while the deformation parameters D_i capture local facial muscle movements that contribute to expression dynamics. The translation vector T_i represents spatial displacement of facial keypoints, ensuring accurate head movement transfer, and the rotation matrix R_i encodes head pose transformations.

Additionally, eye and lip conditions, $C_{\text{eyes},i}$ and $C_{\text{lip},i}$, were extracted to refine expression synthesis. The source and driving implicit keypoints were then transformed accordingly, enabling seamless motion transfer and realistic facial animations from static images. Figure 5 presents the style transfer results on two distinct subjects, using two different sets of driving videos. It can be observed from Figure 5 that our approach effectively captures and transfers a wide range of facial expressions, including subtle micro-expressions and detailed facial muscle movements. The results demonstrate realistic eye and lip dynamics, preserving the natural motion patterns of the driving video. This highlights the effectiveness of our method in achieving high-fidelity expression synthesis and enhancing the realism of animated facial renderings.

C. NEUROMORPHIC EVENT DATA SIMULATION RESULTS

To visually validate the simulated events, they were converted to a frame-like representation. A common approach to do this is called an event histogram [49], which is based on a pixel-wise summation of the polarities of a group of events. Each histogram will act as a frame in a video for qualitative analysis, so the event streams were divided into discrete time-windows to match the frame rate of the original videos. The histogram pixel values were clipped to a range of (-5,5) before normalization to remove outliers that can cause sudden changes in the brightness and visibility of most other events in the histogram. Some example histograms from one event

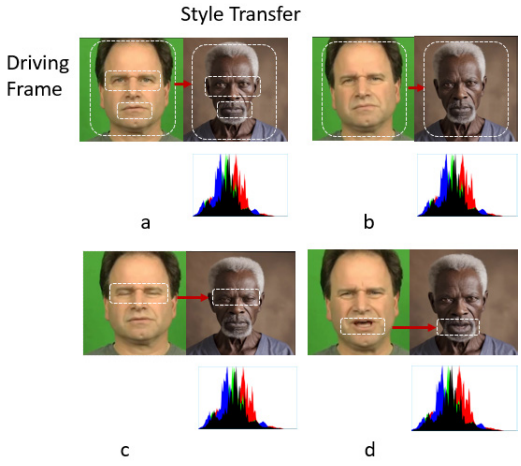


FIGURE 6. Subtle facial muscle movements illustrating micro-expressions with histogram representation.

stream are shown side-by-side with their source frames in Fig. 7. More histograms of events simulated from female samples of ethnicities are shown in Fig. 8, and for male samples in Fig. 9. A qualitative analysis of these events is carried out in section V through the downstream task of facial landmark detection.

D. 2D-3D MORPHING

We leveraged the generated 2D images to construct detailed 3D textures, ensuring a high level of realism and consistency in the final results. To maintain compatibility with standard 3D modeling and rendering software, all 3D models were saved in the .obj file format, which preserves both the geometric structure and texture mapping information. Figure. 10 visually illustrates the transformation process by presenting a side-by-side comparison of the original 2D images and their corresponding 3D results, highlighting the effectiveness of generating high-fidelity textured 3D models.

V. DATA VALIDATION ANALYSIS

A. SYNTHETIC DATA QUALITY EVALUATIONS

To evaluate the quality and realism of our multimodality synthetic data, we employed a range of qualitative and quantitative metrics. These include CLIP score for semantic alignment with text prompts, t-SNE visualization for distributional similarity with real data, BRISQUE for no-reference image quality assessment, and Kernel Inception Distance (KID) for measuring feature-level fidelity. Additionally, we performed identity similarity analysis using cosine similarity to ensure structural consistency and identity preservation across samples.

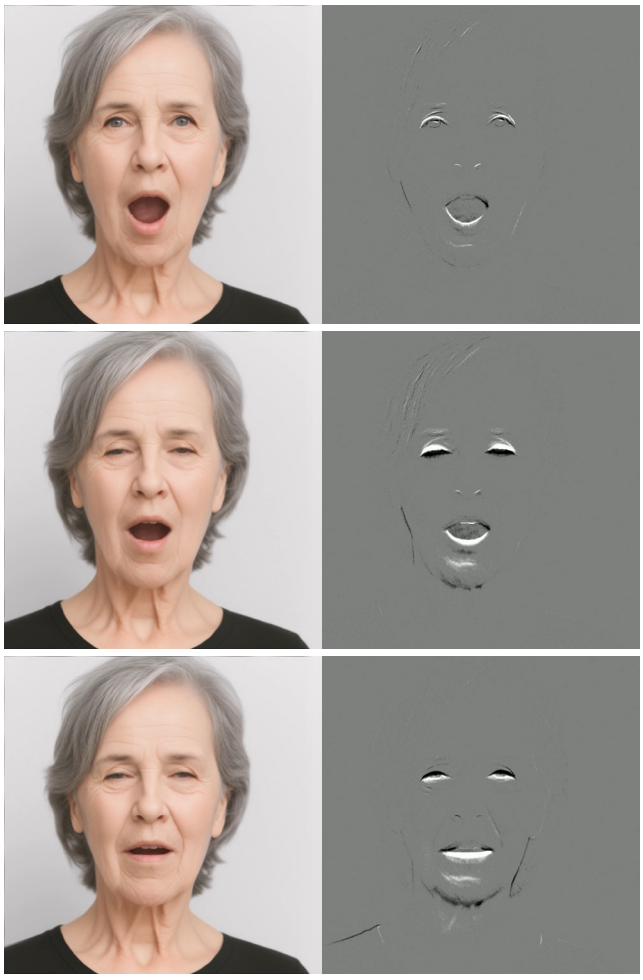


FIGURE 7. Comparison of RGB video frames to their corresponding synthetic events in a histogram representation.

1) CLIP SCORE

To assess the semantic alignment between generated synthetic adult images and their corresponding text prompts, we utilized the CLIP Score. By measuring the cosine similarity between image and text embeddings produced by the CLIP model, this metric helps quantify how well the generated content captures the intended attributes. A higher CLIP score indicates stronger semantic consistency, making it particularly useful for evaluating prompt-guided diffusion outputs. Table 1 shows the results of CLIP Score computed between the synthetic adult images for all the races and their corresponding text prompts. The scores indicate a strong semantic alignment, demonstrating that the generated images effectively capture the attributes described in the prompts, thus validating the efficacy of our prompt-guided generation approach.

TABLE 1. The results of clip scores.

	Female	Male	
		beard	w/o beard
Asian	34.29	35.26	35.97
African	34.44	30.50	32.57
White	31.22	30.47	30.95
Average	33.31	32.63	

2) TSNE VISUALIZATION

To qualitatively assess the distributional similarity between real and synthetic adult data, we employed t-SNE (t-distributed Stochastic Neighbor Embedding) for visualizing high-dimensional image embeddings in a 2D space. As shown in Figure 11, the overlapping clusters of real and synthetic samples indicate that our generated data captures similar feature representations, supporting the fidelity and diversity of the synthetic dataset.

In another experiment, we have compared the distribution similarity of synthetic data samples with real-world data samples. This was done by selecting 100 samples per class from both the genders (male and female) from our dataset. Furthermore, we selected 100 female and 100 male samples from the UTKFace dataset, restricted to individuals with age labels greater than 50 years. We then performed t-SNE using these samples to visualize the data distribution. Figure 11 (b) presents a 3D t-SNE visualization comparing our synthetic dataset with real-world UTKFace samples across gender categories. As shown, the embeddings of synthetic female

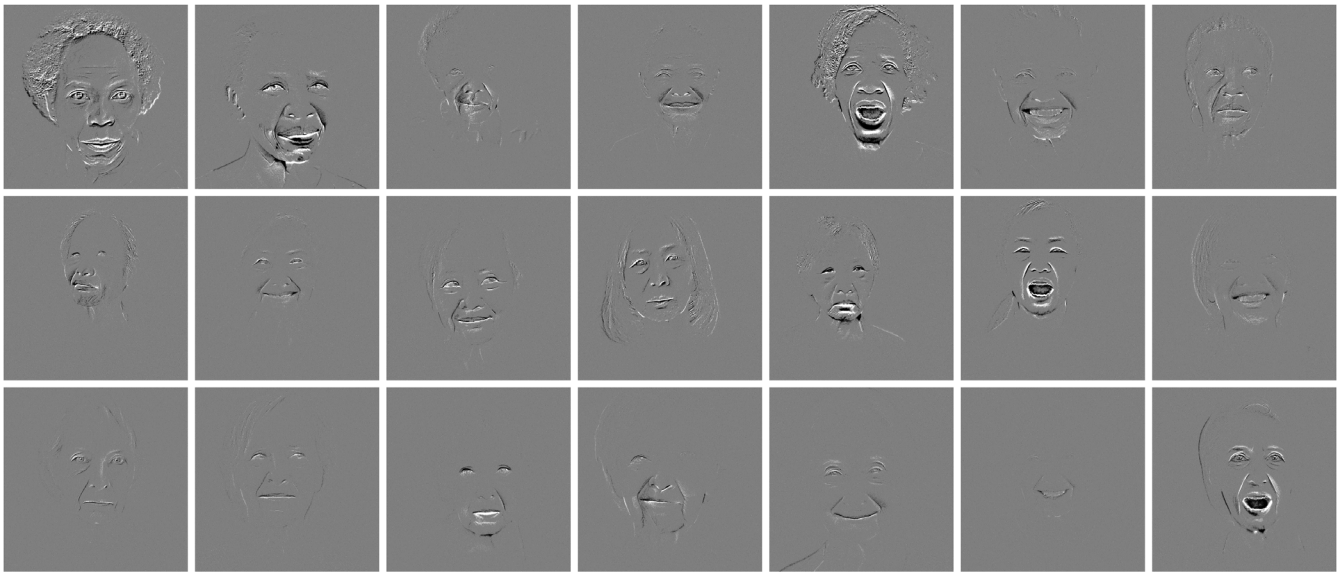


FIGURE 8. Adult female simulated event data visualizations: The first row shows simulations from African samples, the second row Asian samples, and the third row White samples.

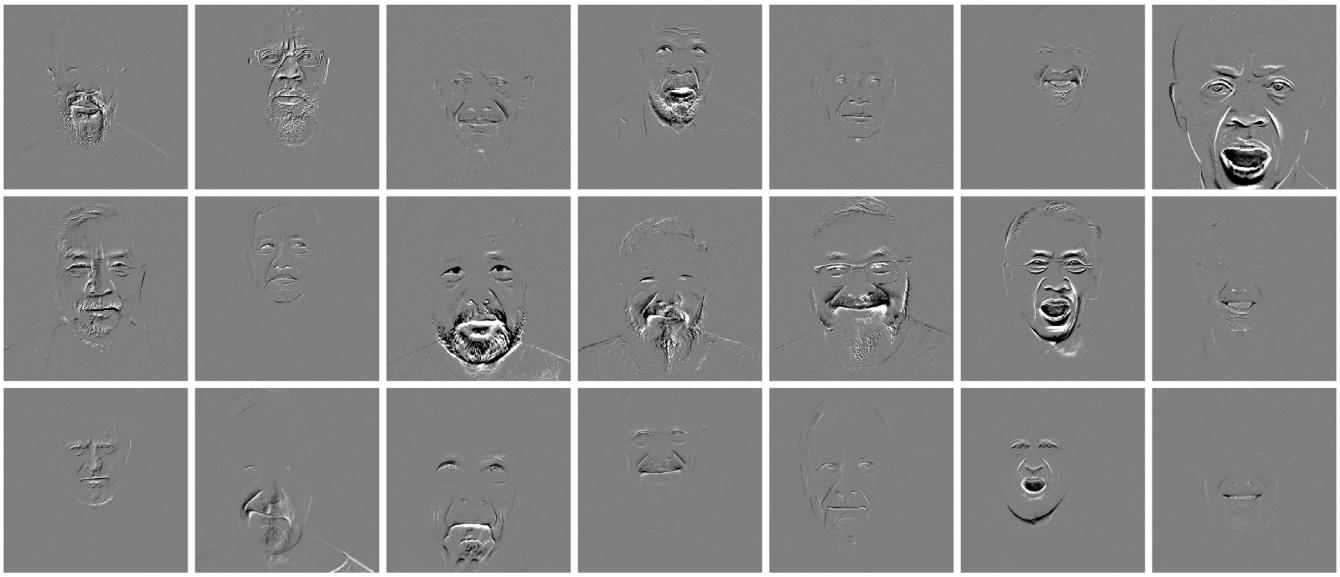


FIGURE 9. Adult male simulated event data visualizations: The first row shows simulations from African samples, the second row Asian samples, and the third row White samples.

and male (in blue and green) are closely clustered with those from Female-UTKFace and Male-UTKFace (in purple and orange), respectively. This strong overlap indicates that our synthetic data preserves meaningful feature distributions and demonstrates high semantic similarity to real-world face data, supporting its utility for downstream tasks like face classification or recognition.

3) KID

Kernel Inception Distance (KID) is a widely used metric for evaluating the quality of synthetic images by comparing their feature distributions to those of real images. Unlike FID, KID is unbiased and better suited for smaller sample sizes. It computes the squared Maximum Mean Discrepancy (MMD) between Inception features of real and generated images, with lower values indicating higher similarity.

Table 2 shows the results of KID computed between the synthetic adult dataset and the corresponding real image distribution. The low KID score suggests that our generated samples are close to the real distribution in the feature space, reinforcing the effectiveness of our generation pipeline.

4) BRISQUE

Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE) is a no-reference metric used to assess the perceptual quality of images by analyzing natural scene

FIGURE 10. 3D texture results: First row show the original facial images, second row shows 3D texture results of these images.

statistics in the spatial domain. It quantifies distortions without the need for a reference image, where lower scores indicate higher perceptual quality and realism. Table 2 shows the results of BRISQUE computed on our synthetic adult dataset, indicating that the generated images exhibit high visual fidelity with minimal artifacts, further validating the effectiveness of our enhancement pipeline.

TABLE 2.

	Female	Male	All
KID	0.080	0.064	0.073
BRISQUE	15.85	9.23	11.45

5) IDENTITY SIMILARITY COMPARISON

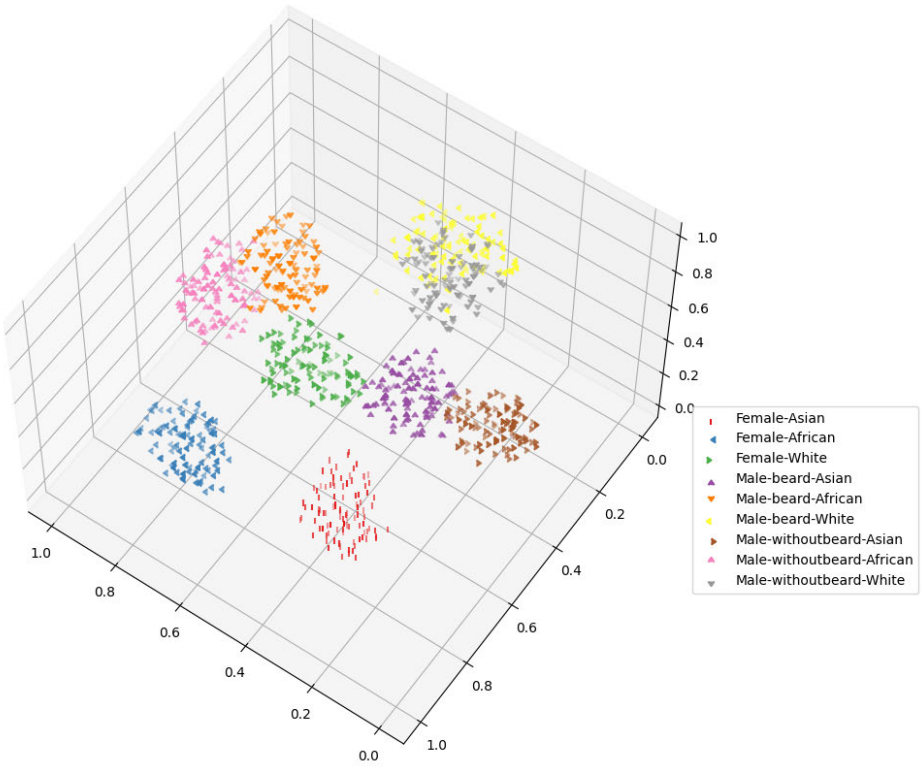
Identity Similarity Comparison evaluates the consistency of identity across synthetic and real faces by comparing their feature embeddings. In our approach, we used cosine similarity on embeddings extracted from a pre-trained facial recognition model. To further assess the accuracy of identity preservation, we utilized the ROC curve to evaluate the true positive rate (TPR) and false positive rate (FPR) for identity matching. A high Area Under the Curve (AUC) indicates strong identity consistency between the synthetic and real faces, confirming the reliability of our synthetic data in preserving identity features. Figure 12 shows the result of ROC curve analysis across three groups: combined, male, and female subjects. The curve for all groups shows a steep ascent toward the top-left corner, indicating near-perfect classification performance. Specifically, the combined and male curves reach a TPR close to 1.0 at extremely low FPR values (as low as 10^{-4}), reflecting strong identity verification capabilities. While the female curve also demonstrates high performance, it shows a slightly lower TPR at the lowest FPR range, suggesting a minor but notable difference in identity matching accuracy across gender. Overall, the ROC curves affirm the high fidelity of identity preservation in our synthetic face generation pipeline.

B. PIPNET FACIAL LANDMARK COMPARISON

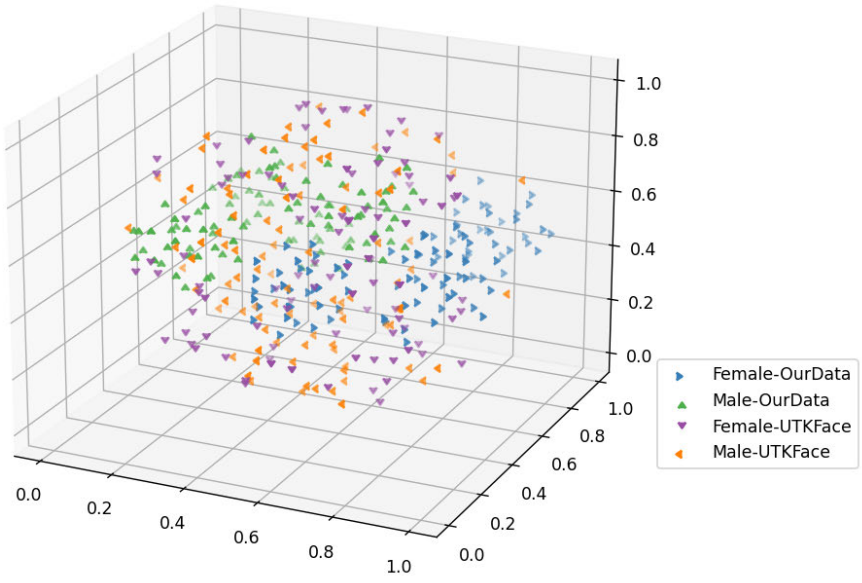
To qualitatively assess the synthetic events, their usability in the downstream task of facial landmark detection was tested. The event-based facial landmark detector from [50] was used as it detects 98 different points with high accuracy. Additionally, this network was adapted from the Pixel-in-Pixel Net (PIPNet) [51] trained on RGB face images, which can be used to obtain reference facial landmarks on the RGB portrait videos. While not comparable to manually annotated ground truth labels, they nonetheless can provide us insight into the relative error between the modalities. The event network was trained on face crops that tend to fit more tightly around the face than those output by the RGB network, especially around the brow and chin. Instead of using the RGB bounding box for cropping the event input, a new face boundary was inferred from the outermost landmarks on each edge followed by a 5% increase to the width and a 10% increase to the height.

Following the procedure from [50], the event streams were divided into non-overlapping time windows with durations of 1/30s that were converted into decaying time-surfaces. This representation is used for its preservation of spatial information across multiple event windows. The time-surface event videos were passed through the landmark network and the resulting landmark positions were saved. Some of these time-surfaces the predicted landmarks are shown in Fig. 13. The normalized mean error (NME), as defined in [51], was calculated between each RGB landmark and its event counterpart. This error was calculated for every frame except the first 5 of every video to allow for adequate initialization of the time-surfaces and LSTM layers for the event network. The average NME for each ethnicity and gender is given in Table 4.

We observe that across all 3 ethnicities, the error on female samples is lower than males. To investigate potential causes, the NME on male samples are split into bearded and non-bearded categories in Table 5. Additionally, instead of using the average error of all landmarks, they are grouped by the facial features they represent. These landmark groups can are colour-coded in Fig. 13.



(a) Compare with different race



(b) Compare with real-world dataset

FIGURE 11. 3D t-SNE Visualization.

These results highlight how the cheek landmarks contribute more to the error than any other group, even more so for males than females and further still for bearded men specifically. The facial features with the largest difference in NME between all male and all female samples are the cheeks, nose, and mouth at 0.95%, 0.72%, and 0.64%, respectively. These are all features that could be partially occluded by a beard so we believe it is possible the landmark network has learned to predict them more confidently on female faces. Some examples of this can be seen in the much higher error

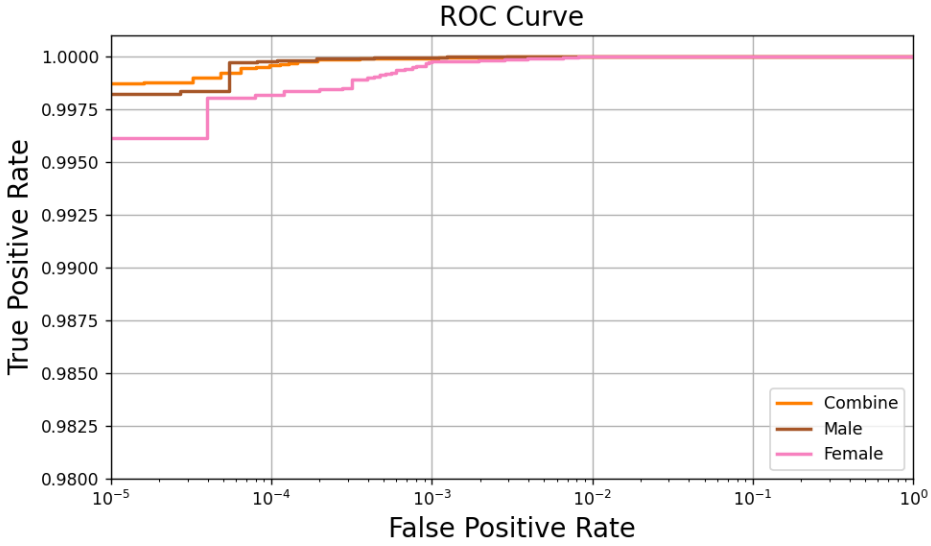


FIGURE 12. ROC curve.

TABLE 3. Facial Expression Classification Result.

DeepFace						
	Female			Male		
	Asian	African	White	Asian	African	White
Angry	32.52%	64.29%	32.65%	72.58%	79.89%	77.20%
Happy	85.81%	71.09%	93.73%	85.67%	91.40%	85.84%
Neutral	64.49%	78.39%	46.94%	67.92%	87.63%	43.92%
Sad	43.57%	21.82%	35.89%	37.19%	18.72%	31.03%
LibreFace						
	Female			Male		
	Asian	African	White	Asian	African	White
Angry	59.20%	81.32%	73.47%	75.07%	97.01%	97.25%
Happy	99.67%	99.99%	99.67%	99.33%	99.99%	99.75%
Neutral	95.17%	93.91%	92.22%	80.05%	94.35%	90.61%
Sad	57.74%	25.99%	38.51%	39.26%	29.28%	36.92%

TABLE 4. NME (%) of landmarks across all genders and ethnicities.

	Female	Male	All
African	3.38	3.72	3.59
Asian	2.99	3.50	3.29
White	2.82	3.64	3.32
All	3.06	3.62	3.40

in the lower landmarks of the bearded faces in Fig. 14. Considering this represents our largest source of error while still being reasonably accurate, we believe the realism of the synthetic events is more than sufficient for use in event camera research.

TABLE 5. NME of landmark groups between female, non-bearded male, and bearded male samples.

Landmark Group	Landmark Count	Female	Male	
			No Beard	Beard
Brow	18	2.86	2.92	2.82
Cheek	33	4.51	4.79	6.17
Eye	18	1.54	1.79	1.76
Mouth	20	2.74	3.17	3.59
Nose	9	1.93	2.52	2.79
All	98	3.06	3.36	3.91

C. DOWNSTREAM FACIAL EXPRESSION CLASSIFICATION

To evaluate the realism and utility of our adult synthetic dataset, we conduct downstream facial expression

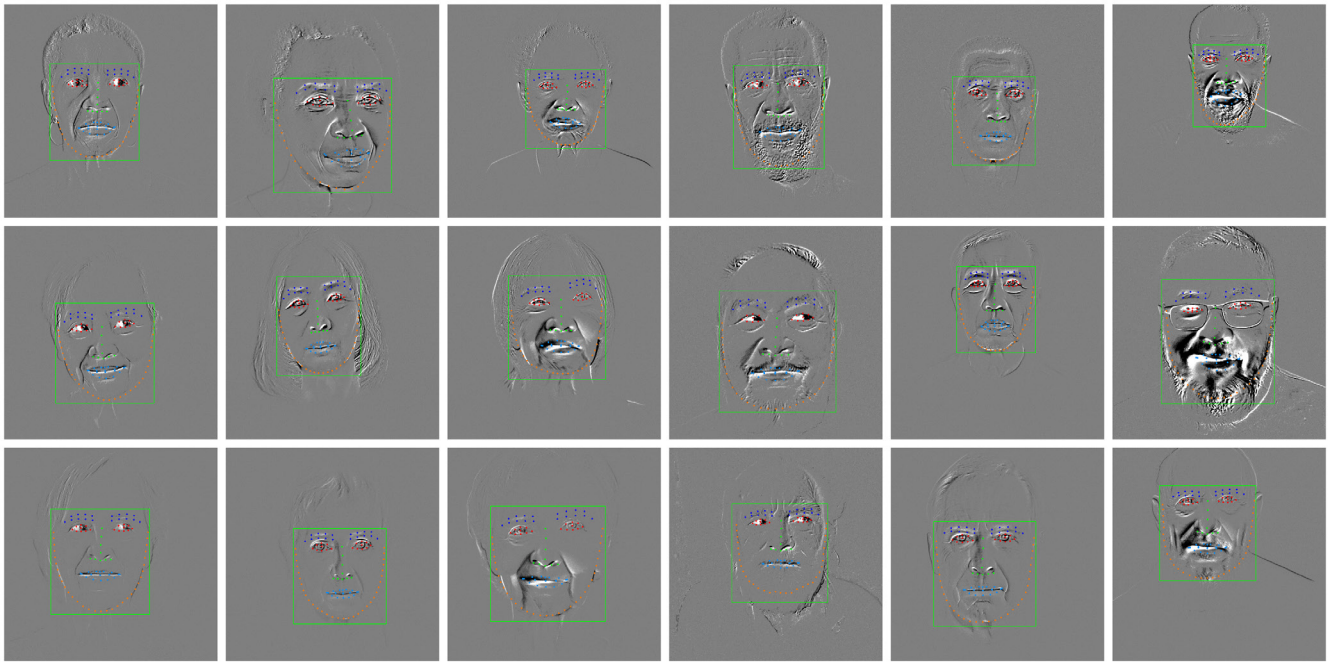


FIGURE 13. Landmark predictions on event time surfaces. The first 3 columns show a female sample from each of the 3 driving videos, repeated with males for the next 3 columns. This is repeated over 3 rows for African, Asian, and White samples respectively. The landmarks are colour-coded by facial feature as follows: Dark Blue = Brow, Orange = Cheek, Red = Eye, Light Blue = Mouth, Green = Nose.

FIGURE 14. Samples of event time-surface videos that were passed through the event landmark network and compared against the RGB source. The driving video and frame number within said video are identical across all 9 samples. Annotated on the time-surfaces are the landmarks detected by the event network, each with a line to its corresponding "ground truth" RGB landmark. The colour of the event annotations transition from green to red at NMEs between 1.5% and 7.5%.

classification using state-of-the-art deep learning models. Specifically, we employ DeepFace [52] and the more recent, open-source LibreFace [53] framework, both of which are widely used in facial analysis tasks. These models are tested

using our synthetic adult face samples as input data, enabling us to assess how well the generated expressions align with established classification boundaries in real-world scenarios. This setup provides a robust benchmark to validate the effectiveness of our dataset in practical computer vision pipelines. Table 3 shows the facial expression classification results across different demographic groups categorized by gender and race (Asian, African, and White). The models were evaluated on their ability to classify four facial expressions: Angry, Happy, Neutral, and Sad. A clear performance disparity can be observed not only between the models but also across demographic subgroups. Overall, LibreFace significantly outperforms DeepFace across nearly all expression categories and demographic segments. The Happy expression yields the highest accuracy in both models, suggesting that it is the most distinguishable facial expression regardless of race or gender. For example, LibreFace achieves nearly perfect accuracy for Happy expressions across all groups, exceeding 99% for most. In contrast, DeepFace, although relatively strong in this category, shows slightly reduced accuracy, especially for African females (71.09%) and Asian females (85.81%).

In both models, the sad expression is the most challenging to classify accurately. DeepFace, in particular, performs poorly for this expression, especially among African and Asian subjects. The accuracy drops as low as 18.72% for African males and 21.82% for African females, indicating substantial bias and difficulty in recognizing sadness in certain racial groups. LibreFace demonstrates some improvement in this regard but still reflects low accuracy for sad expressions in underrepresented groups such as African females (25.99%) and Asian males (39.26%). When comparing racial groups, both models show relatively stronger performance for White and African males, especially in LibreFace, where angry and neutral expression classification often exceeds 90%. However, the performance drops considerably for Asian females across multiple expressions in both models.

In summary, the results highlight critical concerns regarding fairness and bias in facial expression recognition systems. Although LibreFace exhibits improved performance and reduced disparities compared to DeepFace, notable inconsistencies remain particularly in the classification of sad expressions.

VI. SYNTHETIC ADULT MULTIMODALITY DATASET

We introduce a new open-source Synthetic Adult Multimodality Dataset to support research in adult facial analysis and cross-modal representation learning. The dataset includes high-quality 2D facial frames comprising race ethnic dataset and facial expressions, head pose and expression animations, event-based facial data, 2D–3D face morphing, and dense facial landmarks synchronized across modalities. This multimodal composition enables comprehensive exploration of dynamic facial behavior and fusion strategies across spatial and temporal domains. Table 6 summarizes the overall attributes of the dataset. The complete adult multimodality dataset and sample dataset along with the fine-tuned model, is available on our Github Project Webpage: (https://mali-farooq.github.io/SynAdult/).

Table 7 presents a comparative analysis of our proposed SynAdult dataset against other prominent synthetic face datasets, including DigiFace-1M [54], SS-FaceGAN / PFA-GAN [55], and Microsoft FaceSynthetics [56]. While many large-scale face datasets exist in the research community, it is very difficult to find, if not entirely absent, any that offer the same level of multimodal diversity, especially for senior adult data, as provided by SynAdult. The dataset provides 2D RGB images with race diversity, temporal facial video animations with head pose and expressions, corresponding event stream simulations with landmark annotations, adult facial expression, and 3D face representations. This unified and comprehensive setup is uniquely positioned to support emerging biometric applications that span both conventional vision and neuromorphic domains.

VII. DISCUSSION

This work introduces SynAdult, a novel multi-domain synthesis pipeline to render high-quality, photo-realistic adult facial data samples using a text-to-image diffusion model. This pipeline incorporates various modules, including a video retargeting pipeline, a video-to-event representation using the V2E event simulator, and 2D-3D image rendering using UV-IDM models. We analyzed the generated 2D images dataset using three key evaluation metrics: CLIP Score, KID (Kernel Inception Distance), and BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator). The CLIP score results summarized in Table 1 indicates that male subjects without beards achieved the highest alignment (such as 35.97 for Asian males), while female samples maintained consistently strong scores across ethnicities (average of 33.31), confirming the dataset's fidelity to conditioning prompts. The KID and BRISQUE scores, as shown in Table 2, reflect better performance for male images (0.064) compared to female images (0.080), indicating higher realism in male generations. Similarly, the BRISQUE scores, which assess perceptual image quality, revealed lower values for male images (9.23) than female images (15.85), suggesting that male samples generally exhibit sharper and less distorted outputs. Together, these metrics demonstrate that the proposed dataset achieves a strong balance between semantic accuracy and perceptual quality, with slightly better quantitative scores in male samples The video retargeting module demonstrates satisfactory results for both genders by effectively preserving the underlying motion dynamics, such as head pose variations, facial expressions (e.g., smile, frown, surprise), and gaze shifts, while accurately transferring these temporal patterns onto synthetic thermal faces. This ensures that the generated sequences maintain realistic frame-to-frame consistency and natural expression transitions, thereby enhancing the overall animation fidelity across diverse subject categories. The video-to-event representation

TABLE 6. Novel SynAdult dataset attributes.

	2D Data							Video Animations			Event Data			2D-3D Morphing		
	Male		Female	Expressions				Male		Female	Male		Female	Male		Female
	w/b	wob		Anger	Happy	Neutral	Sad	w/b	wob		w/b	wob		w/b	wob	
African	100	100	100	732	726	733	983	10	10	10	10	10	10	100	100	100
White	100	100	100	707	1101	722	950	10	10	10	10	10	10	100	100	100
Asian	100	100	100	687	603	723	865	10	10	10	10	10	10	100	100	100

TABLE 7. Comparison with SoA datasets.

Dataset	Focus	Method	Notes
DigiFace-1M [54]	Face recognition	CG rendering	1M faces, age-varied
SS-FaceGAN / PFA-GAN [55]	Aging progression	StyleGAN variants	Realistic aged faces
Microsoft FaceSynthetics [56]	Face analysis	Diverse synthetic faces	Landmarks & segmentation
SynAdult	Biometric applications and Facial analysis	Stable Diffusion, Video retargeting, V2E, 2D-3D	Races, genders, facial expressions animations

is achieved using the V2E event simulator, which converts standard frame-based videos into event streams by emulating the behavior of neuromorphic event cameras. This process captures pixel-level intensity changes over time, resulting in a temporally rich and low-latency representation of dynamic scenes. By simulating events from video inputs, we enable downstream applications such as motion analysis and event-based recognition. To further assess the quality and alignment of the generated video-to-event representations, we employed a landmark detection network on the simulated event time-surfaces and compared the detected landmarks with the ground truth RGB counterparts. As shown in Fig 14, all nine samples demonstrate consistent driving video frames with corresponding event-based and RGB landmark pairs. The visualizations confirm that the event representations preserve facial geometry across varying demographics, with Normalized Mean Errors (NMEs) ranging from 1.5% to 7.5%. The gradual transition from green to red highlights the spatial deviations between the two modalities, affirming the fidelity of our event simulation pipeline for identity-preserving facial dynamics. Furthermore, we conducted three different downstream machine learning tasks, (1) Identity Similarity Comparison, (2) Facial Landmark Detection/Localization, and (3) Facial Expression Classification. These evaluations go beyond perceptual metrics such as KID, BRISQUE, and CLIP scores, and assess how well the generated data supports real-world tasks relevant to facial analysis and biometrics. These evaluations help demonstrate the dataset's reliability for preserving identity features and capturing subtle expression variations, which are crucial for real-world facial analysis applications.

The proposed approach demonstrates several key strengths that make it a compelling contribution to the field of generative biometrics and synthetic dataset design.

A. KEY STRENGTHS

First, we proposed multimodality synthesis pipeline by the integration of stable diffusion based text-to-image synthesis with video stitching and retargeting framework and further using V2E that allows us to produce temporally consistent, realistic adult facial data that include both static appearance and dynamic motion cues. This is particularly critical for applications such as micro-expression recognition, face anti-spoofing, and low-light biometrics.

Second, the multimodal dataset structure, encompassing both RGB and event data, addresses the need for benchmarking real-time biometric systems under varying conditions (e.g., pose, lighting, occlusion). Our qualitative, quantitative and downstream machine learning analysis shows high fidelity, demographic consistency, and utility of the generated data for real-world biometric applications. Specifically, the synthetic samples demonstrate close alignment with real datasets in terms of visual realism and feature distribution, while also maintaining robustness when used to train and test face expression classification models.

Third, by leveraging few-shot LoRA fine-tuning rather than full retraining, our pipeline achieves efficient model adaptation with minimal computational overhead. This enables rapid generation of tailored datasets for specific demographic subgroups or biometric tasks while remaining resource-efficient.

B. LIMITATIONS AND CHALLENGES

Despite its strengths, the proposed pipeline does have certain limitations. Firstly, one notable limitation is that diffusion models occasionally generate noisy or semantically inconsistent samples, particularly under certain complex prompt guidance conditions. These results are demonstrated in Fig 15 which shows four representative failure cases from our adult diffusion model. These samples highlight common issues like partial face coverage and noisy visual artifacts, especially when using complex or layered textual prompts. As a result, a manual visual inspection and post hoc cleaning process is required to filter out suboptimal outputs and retain only the most realistic and representative data samples for downstream use.

While the diffusion models generate realistic RGB images, the fidelity of simulated event streams is constrained by the

V2E simulator's abstraction assumptions (such as lack of sensor noise and fixed contrast threshold). This may lead to discrepancies in dynamic representation when compared with real neuromorphic camera data.

Moreover, the dataset diversity in terms of ethnic, age, and lighting variation can be further enhanced by introducing more controllable latent conditioning mechanisms in diffusion models and by using style-mixing with adversarial feedback.

FIGURE 15. Failure cases from our SynAdult diffusion model. The top-left image exhibits realistic texture but lacks precise facial symmetry. The top-right sample features a face cropped too tightly, cutting off part of the forehead. The bottom-left image introduces unnatural skin artifacts and inconsistent lighting. The bottom-right image, although aged realistically, fails to capture the full facial geometry due to occlusion and incomplete rendering.

C. BIAS AND FAIRNESS CONSIDERATIONS IN SYNTHETIC SENIOR FACE GENERATION

Given the synthetic generation of senior adult faces, it is crucial to consider the potential biases and fairness implications that may arise. Synthetic datasets, if not properly controlled for demographic balance, can inadvertently reinforce existing societal or algorithmic biases. This is particularly relevant in facial data, where disparities in age, gender, ethnicity, and skin tone representation can affect downstream tasks such as identity verification, emotion recognition, and medical assessments. In our current work, efforts were made to ensure a diverse representation in the conditioning inputs, including a mix of ethnicities and gender-balanced prompts during generation. However, we recognize that biases can still emerge due to the limitations of the training models and latent imbalances in the source data. For instance, diffusion-based models may overfit to overrepresented facial features unless explicitly guided for fairness-aware generation. To address this, we plan to incorporate fairness evaluation metrics in future iterations of the dataset, including demographic parity checks and subgroup performance analysis in downstream tasks. Additionally, we aim to explore techniques such as fairness-aware latent manipulation, diversity-promoting loss functions, and prompt-engineered sampling to enhance demographic coverage and reduce representational bias. Ethically grounded development and validation of synthetic datasets are essential to avoid perpetuating harmful stereotypes or degrading performance for underrepresented groups. By proactively addressing these concerns, we seek to ensure that the synthetic senior face dataset remains both technically robust and socially responsible.

VIII. CONCLUSION AND FUTURE WORK

In this work, we present, a novel pipeline for generating large-scale, multimodal synthetic adult face datasets using diffusion models and event-based simulation. Our framework integrates high-fidelity image generation with neuromorphic event data rendering and 2D–3D face morphing, providing rich, temporally and spatially aligned facial representations. To further enhance realism, we incorporate a video retargeting pipeline for dynamic facial animation and head pose synthesis tailored to adult subjects.

A key strength of our dataset lies in its demographic diversity where we simulate and render data across three major ethnic groups: Asian, African, and White ensuring balanced representation critical for fairness-aware biometric applications. Comprehensive evaluations using KID, BRISQUE, identity similarity, and CLIP scores demonstrate strong fidelity, diversity, and downstream utility. Our synthetic data also shows robust performance in facial attribute classification tasks using SoA models such as DeepFace and LibreFace.

While tuned models occasionally produce noisy outputs requiring manual filtering, and variability exists across prompt complexity, our pipeline remains scalable, memory-efficient, and adaptable. Through this work, we open-source the first multi-race, multimodal synthetic adult face dataset, enabling ethically grounded, fair, and high-resolution data synthesis for critical biometric and computer vision tasks in surveillance, healthcare, and identity verification domains.

Future work will explore: (1) the development of a unified synthetic data generation pipeline capable of spanning all age groups, including children, adults, and seniors, to ensure demographic completeness; (2) domain adaptation strategies to further reduce the sim-to-real performance gap; and (3) incorporation of human-in-the-loop validation to assess and improve identity preservation and expression realism. Additionally, extending the dataset to encompass other biometric modalities such as gait, iris, and hand geometry presents a valuable direction for broadening its applicability in multimodal biometric systems.

ACKNOWLEDGMENT

The authors would like to thank the developers and the contributors of Stable Diffusion and related open-source

generative models for making their work publicly available. Their efforts in advancing AI-driven image synthesis, which have been invaluable to the progress of this research. They also recognize the broader open-source community for its commitment to innovation and for providing valuable resources that support research in generative AI.

REFERENCES

[1] F. X. Gaya-Morey, J. M. Buades-Rubio, P. Palanque, R. Lacuesta, and C. Manresa-Yee, "Deep learning-based facial expression recognition for the elderly: A systematic review," 2025, *arXiv:2502.02618*.

[2] J. Huo, Y. Yu, W. Lin, A. Hu, and C. Wu, "Application of AI in multilevel pain assessment using facial images: Systematic review and meta-analysis," *J. Med. Internet Res.*, vol. 26, Apr. 2024, Art. no. e51250.

[3] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, and H. Zhou, "A survey on computer vision techniques for detecting facial features towards the early diagnosis of mild cognitive impairment in the elderly," *Syst. Sci. Control Eng.*, vol. 7, no. 1, pp. 252–263, Jan. 2019.

[4] E. B. Sönmez, "A computational study on aging effect for facial expression recognition," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 27, no. 4, pp. 2430–2443, Jul. 2019.

[5] H. Ko, K. Kim, M. Bae, M.-G. Seo, G. Nam, S. Park, S. Park, J. Ihm, and J.-Y. Lee, "Changes in computer-analyzed facial expressions with age," *Sensors*, vol. 21, no. 14, p. 4858, Jul. 2021.

[6] J. J. O. Bonet, "Study on the effects of face aging in facial expression recognition," in *Proc. World Congress Eng.*, London, U.K., 2016, pp. 465–470.

[7] M. A. Farooq, W. Yao, G. Costache, and P. Corcoran, "ChildGAN: Large scale synthetic child facial data using domain adaptation in StyleGAN," *IEEE Access*, vol. 11, pp. 108775–108791, 2023.

[8] M. A. Farooq, D. Bigioi, R. Jain, W. Yao, M. Yiwere, and P. Corcoran, "Synthetic speaking children—Why we need them and how to make them," in *Proc. Int. Conf. Speech Technol. Hum.-Comput. Dialogue (SpeD)*, Oct. 2023, pp. 36–41.

[9] W. Yao, M. Ali Farooq, J. Lemley, and P. Corcoran, "Synthetic face ageing: Evaluation, analysis and facilitation of age-robust facial recognition algorithms," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 7, no. 3, pp. 471–483, Jul. 2025.

[10] M. A. Farooq, W. Yao, and P. Corcoran, "ChildDiffusion: Unlocking the potential of generative AI and controllable augmentations for child facial data using stable diffusion and large language models," *IEEE Access*, vol. 13, pp. 96616–96634, 2025.

[11] Z. Liao, Q. Xie, C. Chen, H. Lu, and Z. Deng, "FaceScore: Benchmarking and enhancing face quality in human generation," 2024, *arXiv:2406.17100*.

[12] A. Ergasti, C. Ferrari, T. Fontanini, M. Bertozzi, and A. Prati, "Controllable face synthesis with semantic latent diffusion models," 2024, *arXiv:2403.12743*.

[13] S. Banerjee, G. Mittal, A. Joshi, C. Hegde, and N. Memon, "Identity-preserving aging of face images via latent diffusion models," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2023, pp. 1–10.

[14] W. Rowan, P. Huber, N. Pears, and A. Keeling, "Fake it without making it: Conditioned face generation for accurate 3D face reconstruction," 2023, *arXiv:2307.13639*.

[15] Y. Xue, X. Xie, R. Marin, and G. Pons-Moll, "Human-3Diffusion: Realistic avatar creation via explicit 3D consistent diffusion models," 2024, *arXiv:2406.08475*.

[16] X. Di and V. M. Patel, "Multimodal face synthesis from visual attributes," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 3, no. 3, pp. 427–439, Jul. 2021.

[17] N. G. Nair, J. M. J. Valanarasu, and V. M. Patel, "Maxfusion: Plug & play multi-modal generation in text-to-image diffusion models," in *Proc. ECCV*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham, Switzerland: Springer, 2025, pp. 93–110.

[18] Q. Zhao, P. Long, Q. Zhang, D. Qin, H. Liang, L. Zhang, Y. Zhang, J. Yu, and L. Xu, "Media2Face: Co-speech facial animation generation with multi-modality guidance," in *Proc. Special Interest Group Comput. Graph. Interact. Techn. Conf. Conf. Papers*, Jul. 2024, pp. 1–13.

[19] F. Becattini, F. Palai, and A. D. Bimbo, "Understanding human reactions looking at facial microexpressions with an event camera," *IEEE Trans. Ind. Informat.*, vol. 18, no. 12, pp. 9112–9121, Dec. 2022.

[20] D. Kang and D. Kang, "Event camera-based pupil localization: Facilitating training with event-style translation of RGB faces," *IEEE Access*, vol. 11, pp. 142304–142316, 2023.

[21] D. Joubert, A. Marcireau, N. Ralph, A. Jolley, A. V. Schaik, and G. Cohen, "Event camera simulator improvements via characterized parameters," *Frontiers Neurosci.*, vol. 15, Apr. 2021, Art. no. 702765.

[22] Y. Hu, S.-C. Liu, and T. Delbruck, "V2E: From video frames to realistic DVS events," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1312–1321.

[23] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," in *Proc. Conf. robot Learn.*, 2018, pp. 969–982.

[24] A. Ziegler, D. Teigland, J. Tebbe, T. Gossard, and A. Zell, "Real-time event simulation with frame-based cameras," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 11669–11675.

[25] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *Seminal Graphics Papers: Pushing the Boundaries*. New York, NY, USA: Association for Computing Machinery (ACM), 2023, pp. 157–164.

[26] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, Jan. 2022.

[27] R. T. Marriott, S. Romdhani, and L. Chen, "A 3D GAN for improved large-pose facial recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13440–13450.

[28] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[29] R. Slossberg, I. Jubran, and R. Kimmel, "Unsupervised high-fidelity facial texture generation and reconstruction," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 212–229.

[30] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[31] A. Tewari, M. Elgharib, G. Bharaj, F. Bernard, H.-P. Seidel, P. Pérez, M. Zollhöfer, and C. Theobalt, "StyleRig: Rigging StyleGAN for 3D control over portrait images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6141–6150.

[32] H. Bai, D. Kang, H. Zhang, J. Pan, and L. Bao, "FFHQ-UV: Normalized facial UV-texture dataset for 3D face reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 362–371.

[33] Z. Chai, H. Zhang, J. Ren, D. Kang, Z. Xu, X. Zhe, C. Yuan, and L. Bao, "REALY: Rethinking the evaluation of 3D face reconstruction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 74–92.

[34] R. Abdal, P. Zhu, N. J. Mitra, and P. Wonka, "StyleFlow: Attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows," *ACM Trans. Graph.*, vol. 40, no. 3, pp. 1–21, Jun. 2021.

[35] H. Li, Y. Feng, S. Xue, X. Liu, B. Zeng, S. Li, B. Liu, J. Liu, S. Han, and B. Zhang, "UV-IDM: Identity-conditioned latent diffusion model for face UV-texture generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 10585–10595.

[36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[38] C. Zhang, R. Barbano, and B. Jin, "Conditional variational autoencoder for learned image reconstruction," *Computation*, vol. 9, no. 11, p. 114, Oct. 2021.

[39] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22500–22510.

[40] J. Guo, D. Zhang, X. Liu, Z. Zhong, Y. Zhang, P. Wan, and D. Zhang, "LivePortrait: Efficient portrait animation with stitching and retargeting control," 2024, *arXiv:2407.03168*.

[41] S. Woo, S. Debnath, R. Hu, X. Chen, Z. Liu, I. S. Kweon, and S. Xie, "ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16133–16142.
[42] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.
[43] T.-C. Wang, A. Mallya, and M.-Y. Liu, "One-shot free-view neural talking-head synthesis for video conferencing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10039–10049.
[44] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1874–1883.
[45] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, Oct. 2014.
[46] H. Jiang, D. Sun, V. Jampani, M.-H. Yang, E. Learned-Miller, and J. Kautz, "Super SloMo: High quality estimation of multiple intermediate frames for video interpolation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9000–9008.
[47] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *CoRR*, vol. abs/1312.6114, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:216078090
[48] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10674–10685.
[49] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.
[50] P. Kielty, C. Ryan, W. Shariff, J. Lemley, and P. Corcoran, "Event-based multi-task facial landmark and blink detection," *IEEE Access*, vol. 13, pp. 45609–45622, 2025.
[51] H. Jin, S. Liao, and L. Shao, "Pixel-in-pixel net: Towards efficient facial landmark detection in the wild," *Int. J. Comput. Vis.*, vol. 129, no. 12, pp. 3174–3194, Dec. 2021.
[52] S. Serengil and A. Özpınar, "A benchmark of facial recognition pipelines and co-usability performances of modules," *Bilişim Teknolojileri Dergisi*, vol. 17, no. 2, pp. 95–107, Apr. 2024.
[53] D. Chang, Y. Yin, Z. Li, M. Tran, and M. Soleymani, "LibreFace: An open-source toolkit for deep facial expression analysis," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 8190–8200.
[54] G. Bae, M. de La Gorce, T. Baltrušaitis, C. Hewitt, D. Chen, J. Valentin, R. Cipolla, and J. Shen, "DigiFace-1M: 1 million digital face images for face recognition," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3515–3524.
[55] Q. T. M. Pham, J. Yang, and J. Shin, "Semi-supervised FaceGAN for face-age progression and regression with synthesized paired images," *Electronics*, vol. 9, no. 4, p. 603, Apr. 2020.
[56] E. Wood, T. Baltrušaitis, C. Hewitt, S. Dziadzio, T. J. Cashman, and J. Shotton, "Fake it till you make it: Face analysis in the wild using synthetic data alone," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3661–3671.

MUHAMMAD ALI FAROOQ (Senior Member, IEEE) received the B.E. degree in electronic engineering from Iqra University, in 2012, the M.S. degree in electrical control engineering from the National University of Sciences and Technology (NUST) Pakistan, in 2017, and the Ph.D. degree from the National University of Ireland Galway (NUIG), in 2022.

Currently, he is a Research Fellow with the University of Galway and a Machine Learning Research Intern with FotoNation, where his research work is focused on building large scale synthetic datasets for real world applications. He received the prestigious H2020 European Union (EU) Scholarship as part of his Ph.D. research and has contributed to several large-scale projects funded by EU and Enterprise Ireland. His research interests include machine vision, computer vision, generative AI, machine learning, large language models (LLMs), thermal imaging, medical imaging, and sensor fusion.

PAUL KIELTY received the B.E. degree in electronic and computer engineering from the University of Galway, in 2021. He is currently pursuing the joint Ph.D. degree with the University of Galway and the ADAPT SFI Research Centre. His research is focused on deep learning methods with neuromorphic vision, with particular interest in driver monitoring tasks.

WANG YAO received the B.Sc. degree in computer science and technology from Southwest University, China, in 2016, the M.Sc. degree in control engineering from the University of Chinese Academy of Sciences (UCAS), in 2019, and the Ph.D. degree from the University of Galway. She is currently a Postdoctoral Researcher with the University of Galway. She is also an Intern with FotoNation. Her research interest includes computer vision.

PETER CORCORAN (Fellow, IEEE) holds the Personal Chair in electronic engineering with the College of Science and Engineering, University of Galway. He was a Co-Founder of several start-up companies, notably FotoNation. He has over 600 technical publications and patents, over 100 peer-reviewed journal articles, 120 international conference papers; and a co-inventor of more than 300 granted U.S. patents. He is an IEEE Fellow, recognized for his contributions to digital camera technologies, notably in-camera red-eye correction and facial detection. He is a member of the IEEE Consumer Electronics Society, for over 25 years. He is the Founding Editor of *IEEE Consumer Electronics Magazine*.