IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 October 2023, accepted 21 November 2023, date of publication 23 November 2023,
date of current version 29 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3336405

RESEARCH ARTICLE

SegTex: A Large Scale Synthetic Face Dataset for Face Recognition

LAUDWIKA AMBARDI[1], SUNGEUN HONG[2], (Member, IEEE), AND IN KYU PARK[1], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, Inha University, Incheon 22212, South Korea
[2]Department of Immersive Media Engineering, Sungkyunkwan University, Seoul 03063, South Korea

Corresponding author: In Kyu Park (pik@inha.ac.kr)

This work was supported in part by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government through MSIT under Grant 2022R1A4A1033549 and Grant NRF-2019R1A2C1006706, and in part by the Inha University Research Grant.

ABSTRACT Face recognition remains challenged by data limitations in both scale and diversity, coupled with the ethical dilemmas of using images without the subjects' consent. To address these issues, this paper presents the SegTex framework, a cutting-edge method for generating synthetic face datasets by converting Segmentation maps into Textured images. Using the CelebAHQ-Mask dataset for segmentation maps and extracting facial features from the CelebAMask-HQ dataset, the SegTex method efficiently creates varied synthetic facial characteristics. This approach not only sidesteps the need for real-world data collection but also offers a rich and diverse dataset, essential for improving face recognition algorithm performance. In our experiments, models trained on the SegTex-generated dataset displayed superior performance metrics when compared to those trained on conventional datasets, underscoring the practical utility of our method. This robust performance, combined with the ethical advantages of synthetic data generation, ensures our approach holds significant importance in the field of face recognition.

INDEX TERMS Face synthesis, synthetic dataset, face recognition.

I. INTRODUCTION

Face recognition technology has become a crucial component in a range of applications, from surveillance systems to biometrics and social media platforms. As its use becomes more widespread, there is an increasing need for systems that are both efficient and ethically grounded. The success of any face recognition system largely depends on the quality and diversity of its training dataset. A robust and representative dataset that captures a broad spectrum of human facial features across different demographic and environmental conditions is essential. Nevertheless, building such a dataset requires a rigorous process, which involves collecting data from varied sources and accurate labeling.

Conventional data collection methods for face recognition often rely on web-based sources. These methods gather vast amounts of images from websites, social media, and other online platforms. While this approach collects diverse images quickly, enhancing face recognition system performance, it raises ethical concerns. Collecting images without the individuals' knowledge or consent has implications beyond individual privacy. It can lead to potential misuse in surveillance or introduce bias into algorithms due to non-representative data. Recognizing these issues, some studies, referenced in [1] and [2], underline the importance of obtaining clear permissions when using publicly accessible images. This situation emphasizes the need for data collection methods that balance both efficiency and ethical considerations [3].

In this context, synthetic data emerges as a notable solution, offering an alternative to traditional data sources. This method revolves around crafting artificial data that replicates the intricacies and variety found in real-world data, without utilizing actual images of individuals. Despite significant advancements in synthetic data generation as shown in Figure 1, they often hinge on pre-existing datasets or demand excessive computational resources to generate a considerable number of images, as highlighted in previous works [4] and [5]. These issues limit their scalability and practicality. In other words, ensuring that synthetic facial data

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

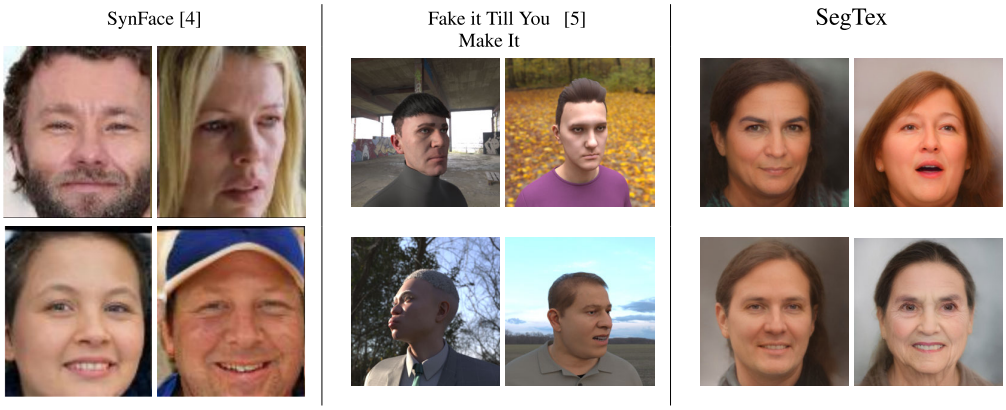


FIGURE 1. A comparison of our proposed synthetic dataset with prior works utilizing synthetic data. While previous studies [4] predominantly depend on existing datasets or require substantial computational resources to produce numerous images [5], our approach stands out. We efficiently generate a vast number of images using moderate computational power, requiring only segmented textures as prior knowledge. Additionally, our method upholds data privacy.

is realistic, unbiased, and representative poses its own set of challenges.

In this study, we address the critical challenge of devising a synthetic dataset for face recognition that prioritizes both privacy and diversity. Understanding the complexities associated with synthetic data, such as potential lack of realism or inadvertent biases due to algorithmic constraints, we have pursued a distinctive synthesis approach. Our approach is anchored by a combination of computational efficiency and an innovative face-generation technique, allowing the swift creation of diverse facial representations. By employing a segmentation map and five textured images, we synthesize faces with meticulously controlled geometry and texture. This approach, reinforced by the fusion of the segmentation map with textured regions, ensures the generation of lifelike synthetic faces, presenting a naturally varied dataset ready for face recognition tasks.

Our methodology goes further by addressing a principal challenge: the seamless integration of textures into a segmentation map. We move away from the traditional approach of using one-hot encoded labels. Instead, we dive deep into understanding the foundational statistics and essence of the segmentation mask. This approach allows our model to create a large number of images, specifically designed for face recognition datasets. Our method not only produces realistic and fair synthetic data but also avoids the ethical issues commonly faced with traditional data collection.

Our proposed method represents a major advancement in face recognition research. It naturally incorporates ethical considerations into the main data collection and processing steps and provides a flexible basis for future research in various face recognition applications. The main contributions of our study are:

- We introduce a synthesis mechanism for a privacy-centric synthetic face dataset. This approach addresses the ethical concerns associated with traditional data collection from web sources.
- Our method integrates textures into a segmentation map, moving away from one-hot encoded labels. Utilizing a segmentation map with five textured images, we generate diverse synthetic faces ideal for recognition tasks.
- We developed a privacy-centric dataset for face recognition, capable of producing 1.8 million images in 48 hours with 30 thousand unique attribute combinations. We extensively validate the utility of our dataset through both quantitative and qualitative analyses.

II. RELATED WORK

A. IMAGE GENERATION TECHNIQUES

Generative adversarial networks (GANs), introduced by Goodfellow et al. [6], changed the way we train models to create images. These networks have been applied to tasks such as creating new images, enhancing image resolution, and altering the style of images [7], [8], [9]. Various methods emerged for converting one type of image into another. Examples include Pix2Pix [10], which changes images using cGANs; CycleGAN [9], which can operate without paired examples; and SPADE [11], which results in higher-quality images. Recent advancements have focused on generating diverse images [12], facilitating multi-domain image transitions [13], and preserving identity during transformations [14].

In this study, while we build upon the foundation of these image-changing techniques, our approach differentiates itself by presenting a novel method to synthesize detailed face images using segmentation maps and textures. This refined method demonstrates our distinct contribution to the field of face image creation.

B. GENERATION OF FACIAL FEATURES

The synthesis of facial features has gained a lot of attention because of its use in face recognition, animation, and virtual reality. Many studies have focused on creating specific parts of the face, such as hair [15], eyes [16], and mouth [17].

Using GANs, these studies have shown success in making realistic facial features while keeping the original identity of the subject. Some other works, like Fader Networks [18] and the Attribute2Image framework [19], have looked at creating full-face images.

Different from these methods, our study uses detailed information from segmentation maps and careful data positioning during preprocessing. This method lets us create and adjust specific features in the images more accurately. This approach provides more flexibility and accuracy in creating images compared to existing methods.

C. FACE RECOGNITION AND SYNTHETIC DATA

Face recognition has been significantly improved by deep learning methods such as DeepFace [20], FaceNet [21], and ArcFace [22]. These methods rely heavily on large datasets of face images. To further increase the diversity of these datasets, Generative Adversarial Networks (GANs) have been used. A notable direction in recent research is the attribute disentanglement technique, like what's seen in the DR-GAN framework [23], which can improve face recognition results. In parallel, there's a growing interest in using synthetic data to further push the boundaries of face recognition, both in terms of evaluating models and creating novel datasets [24], [25]. Our research builds on these advancements. We incorporate the leading face recognition techniques to create and train on our synthetic dataset, aiming to make it widely applicable in various scenarios.

Segmentation maps in image synthesis have led to much finer control over generated images. Works like CRN [26], pix2pixHD [27], and SPADE [11] highlight their effectiveness in this domain. In the realm of face synthesis and editing, techniques like the one proposed by Gu et al. [28] emphasize the importance of these maps in achieving detailed editing and realistic face generation. Moreover, the introduction of techniques like collaborative diffusion [29] has opened doors for leveraging segmentation maps alongside textures, enhancing multi-modal control in generative tasks. Our model, named SegTex, combines the principles from GANs, image-to-image translation, and face synthesis to craft distinct segmentation masks. We then transform these masks into geometric shapes and merge them with textures. Through this, we achieve a wide variety of image generation while preserving the structural consistency, providing a new angle to address face recognition tasks.

III. METHODOLOGY FOR DATASET GENERATION

Our image synthesis methodology primarily revolves around two central processes: the generation of synthetic segmentation maps and the texture application to these synthetic outlines. Initially, we focus on the creation of synthetic segmentation maps. By leveraging Adaptive Instance Normalization (AdaIN) with data from the CelebAMask-HQ dataset, these maps highlight critical regions in images, such as distinguishing facial features. The prime aim here is to clearly outline and differentiate areas of an image, ensuring proper structure within the resultant synthetic depiction. Once the segmentation map is established, our endeavor shifts to enrich these outlines by infusing them with texture. This intricate task is managed by our Texture Infusion Network, denoted as SegTex. The SegTex network processes the segmentation map and other requisite inputs, layering realistic textures onto the marked segments. In essence, our methodology integrates segmentation and texture application to produce high-quality synthetic images.

A. SEGMENTATION MAP GENERATOR

Our dataset creation begins with the development of synthetic segmentation maps. These maps give a structured view of distinct image areas and are essential to our image synthesis method. We separate the segmentation masks from the CelebAMask-HQ dataset into three segments: skin, hair, and mouth. Each segment is then merged to create a unified representation, while keeping each category's distinct features.

The Adaptive Instance Normalization (AdaIN) method is used in our synthesis framework. AdaIN is crucial for effectively merging the feature maps from the segments. The input x denotes the skin feature map, whereas the reference style y is formed from the combined feature maps of hair and mouth. Through AdaIN, the mean and variance of x are aligned to those of y, leading to a smooth combination of features.

$$\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y) \tag{1}$$

This fusion, enabled by the AdaIN layer, produces images that show realistic facial attributes in line with the original segmentation maps.

During training, we use three main loss functions: L1 loss, perceptual loss, and adversarial loss. The L1 loss measures pixel differences between the created and target images. The perceptual loss checks for variations in advanced features, and the adversarial loss, using Binary Cross Entropy (BCE), measures the discriminator's ability to differentiate the synthesized images. The total loss is a combination of the following individual losses.

1) L1 LOSS

The L1 loss, also known as the mean absolute error, captures the absolute difference between the reconstructed and target images:

$$L_{L1} = \text{mean}(|\text{recon} - \text{target}|), \tag{2}$$

pushing for pixel-wise consistency between the regenerated and actual images.

2) PERCEPTUAL LOSS

This loss evaluates the distinction in high-level content and structural information between the reconstructed and target

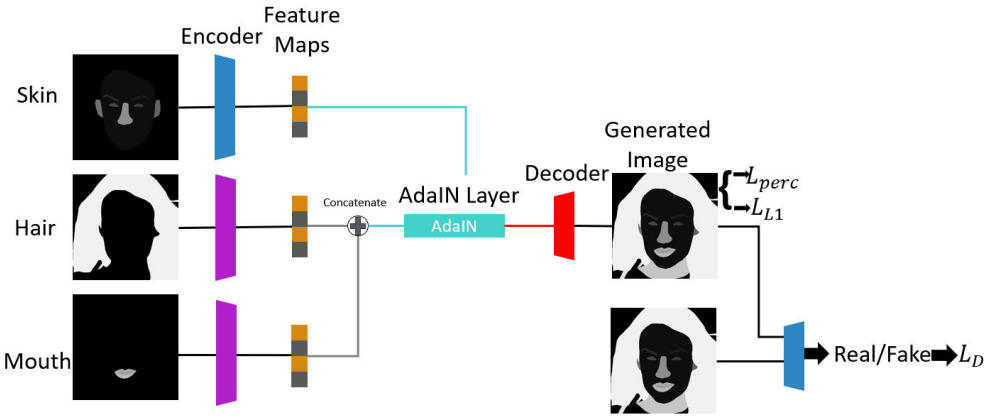


FIGURE 2. Outline of the synthetic segmentation map creation. Distinct regions of the segmentation map–skin, hair, and mouth–are processed by individual encoders, producing specialized feature maps. Concatenation fuses the hair and mouth feature maps, which, when integrated with the skin feature map via the AdaIN operation, results in a unified feature depiction. The synthesized feature map then undergoes decoding to produce the final synthetic segmentation map.

images. It can be mathematically expressed as:

$$L_{\text{perceptual}} = \frac{1}{W_i H_i} \sum_{w=1}^{W_i} \sum_{h=1}^{H_i} (\phi_i(\text{recon})_{w,h} - \phi_i(\text{target})_{w,h})^2 \tag{3}$$

In this equation, $\phi_i(\cdot)$ indicates the feature map activations from the i^{th} layer of a pre-trained VGG network. The terms W_i and H_i denote the width and height of these feature maps, respectively. This loss measures the Mean Squared Error (MSE) between the feature activations of the regenerated and target images, ensuring both have similar high-level features.

3) ADVERSARIAL LOSS

This loss originates from the GAN framework, ensuring that the synthesized images are indistinguishable from real ones. Mathematically:

$$L_{\text{adv}} = \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{4}$$

Here, $D(x)$ gauges the probability of a real data sample x being genuine. $G(z)$ is the generator's output for a given input noise z, and $D(G(z))$ is the discriminator's estimate of that output being a real image. This loss encourages the generator to produce images that the discriminator believes to be genuine, while the discriminator tries to distinguish between real and fake images.

4) TOTAL LOSS

The overall training objective is a weighted sum of the above loss functions:

$$L_{total} = \lambda_{L1} L_{L1} + \lambda_{perc} L_{perceptual} + \lambda_{adv} L_{adv}, \tag{5}$$

With the AdaIN method, we successfully merge skin, mouth, and hair segmentation masks. This integration produces synthetic images that faithfully represent realistic facial attributes, yielding a diverse and high-quality dataset.

B. TEXTURE INFUSION NETWORK

We present SegTex, a distinctive architecture designed for texture synthesis. SegTex ingests six diverse inputs: a segmentation map, skin, hair, left eye, right eye, and mouth. Each input undergoes its specialized encoder module, and the processed results pass through subsequent fusion layers and a blending layer to generate the final output (see Figure 3).

1) ENCODER MODULES

Comprising two convolutional layers, batch normalization, and a leaky ReLU activation, these modules efficiently extract features from the respective inputs.

2) FUSIONLAYER

Merges features extracted from the segmentation map and the other facial features (excluding the mouth). These combined features are further processed via a convolutional layer and a gating mechanism, which uses a sigmoid activation.

3) MOUTHFUSIONLAYER

Exclusively for integrating mouth-related features. This layer amalgamates the features of the mouth with those of the segmentation, subsequently processed through a 1×1 convolutional layer.

4) BLENDINGLAYER

This layer adeptly merges the outputs from the FusionLayer and MouthFusionLayer.

5) SPADEGenerator

A pivotal component, the Spatially-Adaptive (De)Normalization (SPADE) generator synthesizes the final image using the

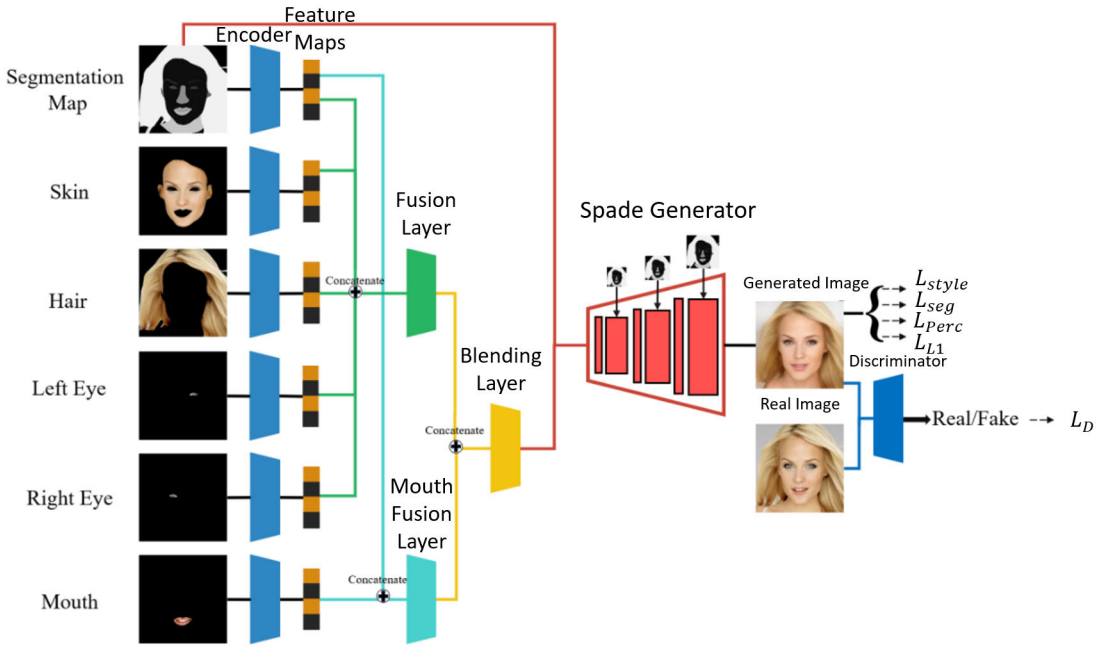


FIGURE 3. Overall flow of the SegTex Framework, aiming at transforming segmentation maps into textured images. The model initiates by processing six distinct images: segmentation map, skin, hair, left eye, right eye, and mouth. Each undergoes encoding, resulting in individualized feature maps. The feature maps from the segmentation map, skin, hair, left eye, and right eye are channeled into a fusion layer for integration. In parallel, the mouth feature map collaborates with the segmentation map feature map in the mouth fusion layer. These outputs are subsequently amalgamated through a blending layer. Finally, the integrated feature map is directed to the SPADE generator, culminating in the generation of the textured image.

fused features and the segmentation map. The operation of SPADE is:

$$\text{SPADE}(x, y) = \gamma(y) \odot x + \beta(y), \tag{6}$$

where x represents the input feature map, y is the segmentation map, and $\gamma(y)$ and $\beta(y)$ are the learned parameters. The $\odot$ represents element-wise multiplication.

During the training phase, a set of loss functions refine the model's outputs, ensuring the synthesis of high-quality, realistic images:

6) STYLE LOSS L_{STYLE}

Assesses the disparity in Gram matrices of the synthesized and target images:

$$L_{style} = \text{mean}((G(\text{recon}) - G(\text{target}))^2), \tag{7}$$

where $G(\cdot)$ calculates the Gram matrix.

7) SEGMENTATION LOSS L_{SEG}

Evaluates differences between the reconstructed and target images, taking into account the segmentation maps:

$$L_{\text{seg}} = \text{mean}\left((\text{recon} \odot \text{seg\_maps} - \text{target} \odot \text{seg\_maps})^2\right). \tag{8}$$

8) MULTISCALE DISCRIMINATOR LOSS L_D

Promotes the synthesis of images that appear realistic across varying scales, using multiple discriminators:

$$L_D = \sum_{i=1}^{N} \left[\frac{1}{2}\text{mean}\left((D_i(\text{recon}) - 1)^2\right) + \frac{1}{2}\text{mean}\left(D_i(\text{target})^2\right) + \lambda_{fm} L_{fm}^{(i)} \right], \tag{9}$$

with N signifying the number of scales, and $L_{fm}^{(i)}$ being the feature matching loss.

9) TOTAL LOSS L_{TOTAL}

The complete objective, represented as a weighted sum of the individual losses:

$$L_{total} = \lambda_{style} L_{style} + \lambda_{seg} L_{seg} + \lambda_{perc} L_{perc} + \lambda_{L1} L_{L1} + \lambda_D L_D, \tag{10}$$

where the weights govern each loss term's influence in the overall optimization.

C. DATASET GENERATION PROCESS

Our dataset creation process is illustrated in Figure 4. It's a comprehensive approach designed for both quality and diversity. Through seven detailed steps, we utilize various techniques, from intricate pose modifications to detailed age adjustments. As a result, we produce a dataset that is not only robust and varied but also closely mirrors real-life human

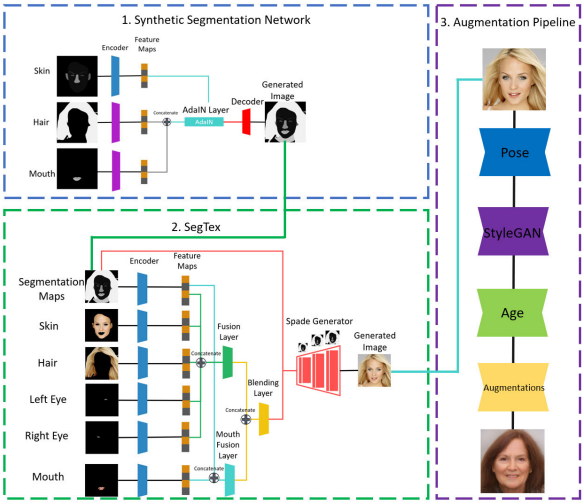


FIGURE 4. The procedure of our dataset generation during the inference phase. Beginning with the synthetic segmentation network, randomized hair and mouth images are utilized to craft a synthetic segmentation map. This map is then augmented with random textures from the CelebAMask-HQ dataset using SegTex, resulting in a detailed textured representation.

attributes, thereby greatly optimizing the performance of face recognition models. This approach ensures a robust, diverse, and realistic dataset, showcasing our commitment to capturing the complexities of human faces, enhancing the performance of face recognition models.

1) SEGMENTATION MAPS GENERATION

Segmentation maps are the foundation of our synthetic image pipeline. For every identity, we produce ten distinct maps, each highlighting different hair and mouth features but maintaining a consistent skin texture. This step ensures the creation of varied facial structures, adding depth to our synthetic dataset.

2) FEATURE RANDOMIZATION

With the segmentation maps ready, we proceed to randomize the associated textures' features, covering skin, hair, eyes, and mouth. This enhancement introduces a new level of diversity to individual facial attributes.

3) APPLYING TEXTURES TO MAPS

In this stage, we overlay the randomized textures onto the generated segmentation maps. The combination of these two elements results in diverse and realistic synthetic images.

4) POSE ADJUSTMENTS FOR TEXTURED IMAGES

This step emphasizes altering the pose of the textured images. Using the facevid2vid [30] method, we craft 20 unique poses for each image, showcasing a plethora of head orientations and facial expressions, enriching the dataset's versatility.

5) INTEGRATION OF THE StyleGAN INVERTER

The pipeline benefits from the inclusion of a StyleGAN inverter, set at a truncation value of 0.6. This tool amplifies the visual fidelity and realism of the images while tapping into StyleGAN's generative power, further diversifying the dataset.

TABLE 1. Model architecture of our encoder to generate segmentation maps.

Stage	Input Size	Output Size
Encoder:skin	[1, 3, 256, 256]	[1, 256, 64, 64]
Encoder:hair	[1, 3, 256, 256]	[1, 256, 64, 64]
Encoder:mouth	[1, 3, 256, 256]	[1, 512, 64, 64]

6) AGE ADJUSTMENTS

A crucial step in the process is the use of the Age transfer module through SAM [31]. This technique modifies the perceived age of the synthetic faces, spanning from ages 10 to 80. Incorporating such age diversity ensures a comprehensive age spectrum in the dataset, improving its authenticity and relevance.

7) IMAGE AUGMENTATIONS

Concluding our process, we apply various augmentations using OpenCV. Adjustments in brightness, contrast, and color balance, along with the application of Gaussian blur, ensure the dataset's resilience across varied conditions.

IV. EXPERIMENTAL RESULTS

A. IMPLEMENTATION DETAILS

Our approach relies on a proficient segmentation network grounded in Adaptive Instance Normalization (AdaIN) [32], complemented by a tailored SPADE generator designed for advanced image segmentation endeavors [11]. The subsequent sections provide a detailed exploration of these fundamental components.

1) GENERATOR FOR SEGMENTATION NETWORKS

The AdaIN function is the linchpin of our segmentation network, amalgamating data from the hair, mouth, and skin regions. Specific encoders for hair and mouth convert images into individual features, which are then merged. Concurrently, the skin image is processed via its exclusive encoder. The output size from the skin encoder is notably twice that of the others (refer to Table 1). Subsequently, the AdaIN function fuses the combined features of the hair and mouth with those from the skin encoder. This forms a composite feature map that incorporates data from all three regions, which is then metamorphosed into a unified image.

2) GENERATOR FOR TEXTURE INFUSION

Our custom SPADE generator, a refined offshoot of the original SPADE model, is tailored for specific image segmentation challenges. It encompasses multiple encoders, each tailored for processing distinct image regions like the skin, hair, eyes (both left and right), mouth, and a comprehensive segmentation map (details in Table 2).

Each encoder processes an RGB image of dimensions (3, 256, 256) and yields a downscaled, feature-extracted

TABLE 2. Model architecture of our encoder and inputs into the modified SPADE generator.

Stage	Input Size	Output Size
Encoder:seg_map	[1, 3, 256, 256]	[1, 256, 64, 64]
Encoder:skin	[1, 3, 256, 256]	[1, 256, 64, 64]
Encoder:hair	[1, 3, 256, 256]	[1, 256, 64, 64]
Encoder:left_eye	[1, 3, 256, 256]	[1, 256, 64, 64]
Encoder:right_eye	[1, 3, 256, 256]	[1, 256, 64, 64]
Encoder:mouth	[1, 3, 256, 256]	[1, 256, 64, 64]
Mouth transform	[1, 256, 64, 64]	[1, 256, 64, 64]
Fusion Layer	[1, 3, 256, 256]	[1, 3, 64, 64]
Mouth Fusion Layer	[1, 256, 64, 64]	[1, 3, 64, 64]
Blending Layer	[1, 3, 64, 64]	[1, 3, 64, 64]

TABLE 3. Main hyperparameter setting.

Hyperparameter	Value
Optimizer	Adam ($\beta_1 = 0.5, \beta_2 = 0.999$)
Learning Rate for G	0.0002
Learning Rate for D	0.0002
Batch Size	32

G: Generator D: Discriminator

version with dimensions (256, 64, 64) using two convolutional layers combined with leaky ReLU activations. The convolution operations maintain kernel size k=4, stride s=2, and padding p=1. Feature integration follows a two-tiered approach. The 'Fusion Layer' consolidates features of the skin, hair, eyes, and segmentation map, while the 'Mouth Fusion Layer' integrates the mouth's features with the segmentation map. Both layers harness concatenation and convolution operations. Following fusion, the 'Blending Layer' evaluates each source's contribution via trainable weights.

A pivotal element is the transformation applied to the mouth encoder's output before its integration through the 'Mouth Fusion Layer'. This includes two convolutional layers, augmented by batch normalization and leaky ReLU activations. In essence, our custom SPADE generator boasts a nuanced design adept at navigating intricate image segmentation challenges. Its structure combines region-centric feature extraction (via encoders), feature consolidation (through fusion layers), and feature selection (via the blending layer). Training hyperparameters can be found in Table 3.

Our SegTex architecture undergoes training from scratch on the CelebAMask-HQ dataset. It comprises a generator and discriminator, leveraging the Adam optimizer with a learning rate of 2×10^{-4} and a batch size of 32. Customizations in the SPADE generator include three classes per channel, differing from the region count in the segmentation map. The network iterates across the dataset for 35 epochs. Both the generator and discriminator compute and back-propagate their losses during each epoch. This training phase spanned approximately two days on an RTX A6000 GPU.

3) DATASET CONSTRUCTION

Our dataset construction heavily relies on a carefully designed data pipeline. We ensure that textures align with specific facial features in images, which optimizes our network's performance. This approach allows us to eliminate the need for more complex layers, like attention layers, resulting in a model that is both lighter and more efficient. Due to the unique variations in the mouth area, such as different shapes and expressions, we incorporated a specialized fusion layer, which significantly improves the mouth's appearance and enhances the overall network's performance.

TABLE 4. Dataset generation comparison.

Method	VRAM used	# of Images	Rendering Time (hours)
Wood *et al.* [5]	1,200GB	100K	48
SegTex (Ours)	96GB	1.8M	48

TABLE 5. Comparison of face recognition datasets.

Dataset	# of People	# of Images	Avg. Images per Class
LFW [33]	5749	13233	2.3
CASIA-WebFace [34]	10575	500k	47.3
MS-Celeb-1M [35]	100k	10M	100
SegTex (ours)	19000	3.8M	200

TABLE 6. Ablation study for each module.

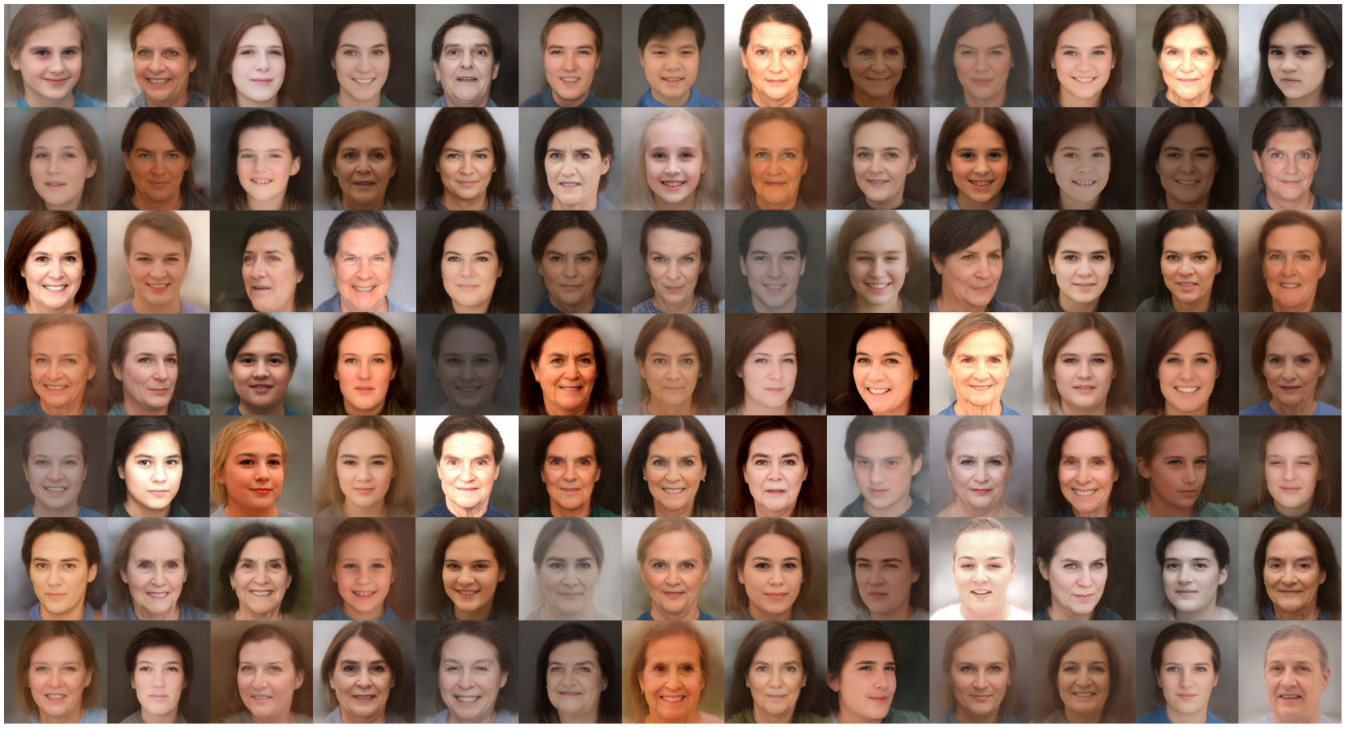
Model	FID Score
SegTex (w/o StyleGAN)	51.57
SegTex (w/ StyleGAN)	113.71
SegTex (w/ StyleGAN and Age Module)	68.49

For our dataset generation, we employ 4 RTX 4090 GPUs, with each GPU running two pipelines. This setup allows us to operate eight pipelines concurrently. Collectively, they utilize about 8.5 GB of VRAM per pipeline, producing approximately 38,400 images every hour or about 920,000 images daily. The numbers, as indicated in Table 4, underscore the effectiveness of our computational framework. Furthermore, our synthetic dataset is expansive and diverse, encompassing 19,000 classes with a total of 3.8 million images. These images have been enriched with various modifications like different poses, ages, and lighting, as detailed in Table 5. Notably, our dataset surpasses its predecessors in size, and by leveraging CelebA-HQ, we guarantee a broad representation of appearances and poses.

B. PERFORMANCE EVALUATION

1) FACE RECOGNITION ASSESSMENTS

To highlight the effectiveness of the synthetic dataset produced by our approach, we conducted a face recognition task. We used 50 images from each class within our synthetic dataset for training, and the remainder were reserved for validation. The training utilized the SoftMax method, with the



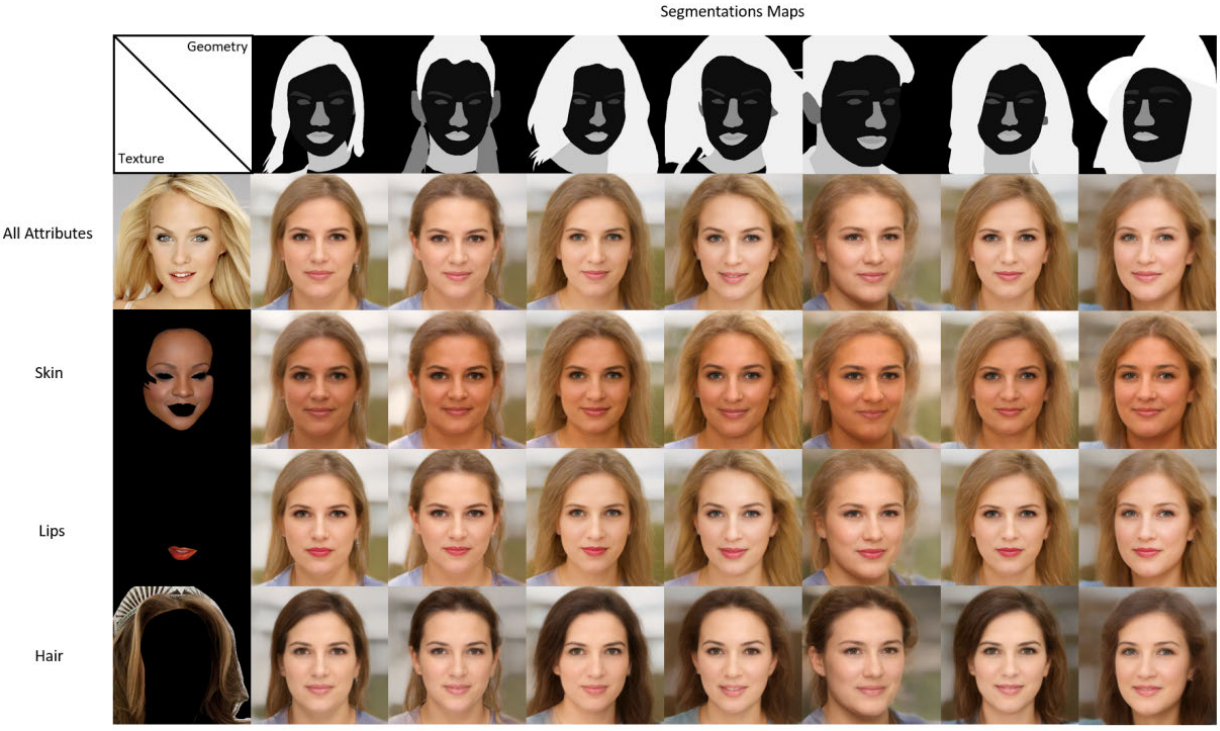
(a)

(b)

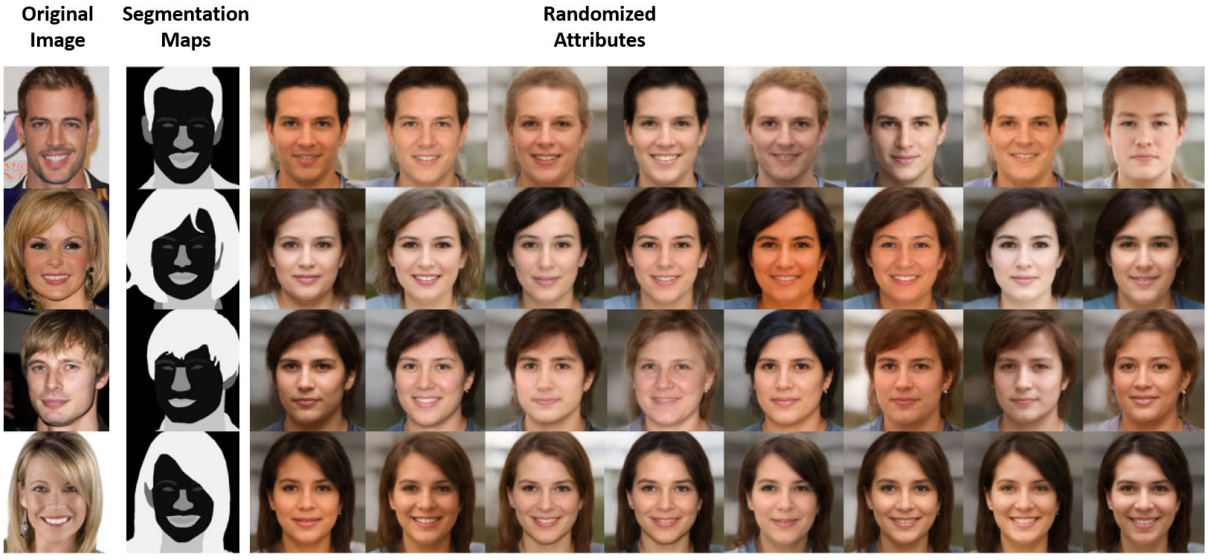
FIGURE 5. Dataset examples using the proposed technique. (a) a selection from our dataset showcasing a wide range of ages, ethnicities, and genders. (b) variations of a single individual, highlighting changes in poses and age progression.

ResNet50 structure serving as the foundation of the model. Throughout the training phase, a conventional cross-entropy loss function was adopted, with a set learning rate of 0.01.

To assess the strengths and areas for enhancement of our SegTex pipeline, we utilize both quantitative face recognition metrics and qualitative evaluations for image realism and



(a)



(b)

FIGURE 6. Qualitative visualization of attribute manipulation. (a) demonstration of SegTex's proficiency in selectively modifying facial attributes while preserving the underlying identity. (b) illustration of generating distinct personalities by randomizing facial attributes.

diversity. Initial evaluations reveal that the standalone SegTex model achieved an FID score of 51.57. However, with the integration of StyleGAN, the FID score was 113.71. Subsequent refinements using the Age Module improved this to 68.49, as detailed in Table 6.

Evaluation results in Table 7 indicate that our SegTex Subset, comprising 7,018 classes, outperforms other datasets by achieving a validation accuracy of 92.04%. Notably, our dataset's superior performance, even with fewer classes compared to CASIA-WebFace, underscores our synthetic data's quality. The enhanced performance observed when combining SegTex with other datasets further attests to the benefits of integrating our synthetic data into face recognition models.

TABLE 7. Performance improvement with the proposed dataset.

Dataset	Accuracy(%)
CASIA-WebFace [22]	77.42
SegTex Subset (ours)	92.04
VGG Face 2 [36]	65.07
SegTex(ours) + CASIA-WebFace [22]	84.32
SegTex(ours) + VGG Face 2 [36]	90.68

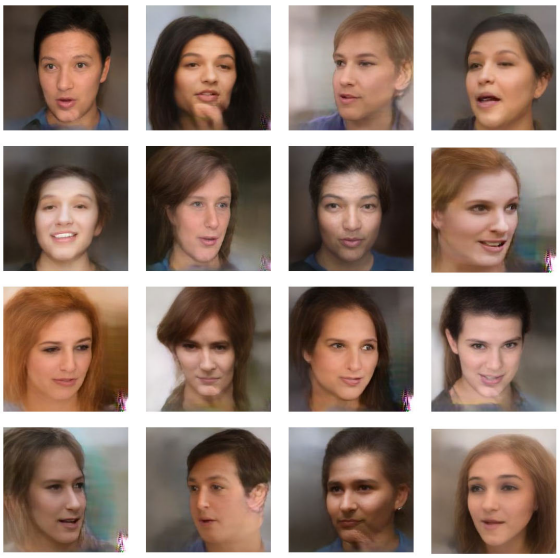


FIGURE 7. Images after pose transfer but before applying a styleGAN inverter to enhance the image. it is visible there are many artifacts if the changes are large.

2) IN-DEPTH ANALYSIS

a: SYNTHESIZED SAMPLE IMAGES

Our model's capability in generating a diverse range of unique and identical individuals is displayed in Figures 5. This highlights the potential for an infinite and varied dataset creation.

b: ATTRIBUTE EDITING

Figure 6 (a) illustrates the flexibility of SegTex in altering facial features by manipulating texture map segments. While enabling minor attribute modifications, the model preserves the original identity, thereby enriching the dataset with variations such as simulated makeup.

c: RANDOM IDENTITY GENERATION

Figure 6 (b) demonstrates the generation of random identities using different textures, emphasizing the model's capacity for vast diversity. Employing the CelebA-HQ dataset, we generate images representing various ethnic groups. It's noteworthy that the focus on European American/White faces in our results is a choice for demonstration and not a model limitation.

d: ARTIFACTS BEFORE INVERSION

The success of our processing pipeline, comprising seven unique modules, hinges on the optimal functioning of each module. Discrepancies in any module can affect the final image quality. Recognizing this, we incorporate a StyleGAN inverter as both an image enhancer and a corrector for potential artifacts. Notably, the most significant artifacts emerge from the pose generator using FaceVid2Vid [30]. Extreme pose changes sometimes produce artifacts beyond the capability of our texture infusion network. The StyleGAN inverter mitigates these issues, ensuring the delivery of pristine, high-quality images.

V. CONCLUSION

In this study, we emphasize the crucial role of synthetic datasets in the field of face recognition. The lack of diverse and extensive real-world data often poses challenges for researchers, and our proposed SegTex framework presents an innovative solution to address these issues. Using segmentation maps from the CelebAHQ-Mask dataset and extracting facial features from the CelebAMask-HQ dataset, SegTex effectively generates a wide range of facial characteristics. By leveraging segmentation maps and various augmentation techniques, we were able to create a dataset with 200 unique images per identity, setting a new standard in synthetic data generation. Our experimental results provide solid evidence for our approach's effectiveness. When tested with a standard face recognition classifier, our dataset demonstrated improved model performance, highlighting its potential in face recognition tasks. As the domain of face recognition continues to grow, it is essential to have innovative dataset creation methods. Our research serves as a foundation for future studies, underlining the importance of high-quality data in advancing face-related tasks.

ACKNOWLEDGMENT

The authors would like to thank Suprema AI Inc., (Kiduk Lee and Dr. Bong Seop Song) for the continued support of the project. The author Laudwika Ambardi would like to thank Prof. Jee Eun Karin Nam and Prof. Eun Ji for their support in helping him navigate the challenges encountered during his graduate study.

REFERENCES

[1] V. U. Prabhu and A. Birhane, "Large image datasets: A pyrrhic win for computer vision?" 2020, *arXiv:2006.16923*.
[2] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 59–68.
[3] Data Society Research Institute. (2019). *Facial Recognition Technologies: A Primer*. [Online]. Available: https://datasociety.net/pubs/ia/Data_Society_Facial_Recognition_Primer_2019.pdf
[4] 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.*, Oct. 2021, pp. 3661–3671.
[5] H. Qiu, B. Yu, D. Gong, Z. Li, W. Liu, and D. Tao, "SynFace: Face recognition with synthetic data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 10860–10870.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 39–44, 2020.
[7] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–16.

[8] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 105–114.
[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2242–2251.
[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5967–5976.
[11] 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.*, Jun. 2019, pp. 2332–2341.
[12] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2018, pp. 4463–4472.
[13] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.
[14] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville, "Augmented CycleGAN: Learning many-to-many mappings from unpaired data," in *Proc. Int. Conf. Mach. Learn.*, Sep. 2018, pp. 452–468.
[15] M.-Y. Liu, X. Huang, A. Mallya, T. Karras, T. Aila, J. Lehtinen, and J. Kautz, "Few-shot unsupervised image-to-image translation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 10550–10559.
[16] B. Dolhansky and C. C. Ferrer, "Eye in-painting with exemplar generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7902–7911.
[17] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "AttGAN: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.
[18] G. Lample, N. Zeghidour, N. Usunier, A. Bordes, L. B. Dinh, and Y. Bengio, "Fader networks: Manipulating images by sliding attributes," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2017, pp. 5969–5978.
[19] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 776–791.
[20] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1701–1708.
[21] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 815–823.
[22] J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 5962–5979, Oct. 2022.
[23] L. Tran, X. Yin, and X. Liu, "Disentangled representation learning GAN for pose-invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1283–1292.
[24] A. Makrushin, A. Uhl, and J. Dittmann, "A survey on synthetic biometrics: Fingerprint, face, iris and vascular patterns," *IEEE Access*, vol. 11, pp. 33887–33899, 2023.
[25] F. Boutros, V. Struc, J. Fierrez, and N. Damer, "Synthetic data for face recognition: Current state and future prospects," *Image Vis. Comput.*, vol. 135, Jul. 2023, Art. no. 104688.
[26] Q. Chen and V. Koltun, "Photographic image synthesis with cascaded refinement networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1520–1529.
[27] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[28] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, "Mask-guided portrait editing with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3431–3440.
[29] Z. Huang, K. C. K. Chan, Y. Jiang, and Z. Liu, "Collaborative diffusion for multi-modal face generation and editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2023, pp. 6080–6090.
[30] 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.*, Jun. 2021, pp. 10034–10044.
[31] Y. Alaluf, O. Patashnik, and D. Cohen-Or, "Only a matter of style: Age transformation using a style-based regression model," *ACM Trans. Graph.*, vol. 40, no. 4, pp. 1–12, Aug. 2021.
[32] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1510–1519.
[33] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces 'Real-Life' Images, Detection, Alignment, Recognit.*, 2007, pp. 1–11.
[34] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," 2014, *arXiv:1411.7923*.
[35] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 87–102.
[36] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2018, pp. 67–74.

LAUDWIKA AMBARDI received the B.Comp.Sc. degree in computer science from Bina Nusantara University, Indonesia, in 2020, and the M.S. degree from Inha University, South Korea, in 2023. He is currently a Computer Vision Researcher with Suprema, South Korea. His current research interests include face synthesis, face recognition, and generative models.

SUNGEUN HONG (Member, IEEE) received the B.S. degree in computer engineering from Hanyang University, South Korea, in 2010, and the M.S. and Ph.D. degrees in computer science from the Korea Advanced Institute of Science and Technology (KAIST), in 2012 and 2018, respectively. He is currently an Assistant Professor with Sungkyunkwan University, South Korea. Prior to his position at Sungkyunkwan University, he was an Assistant Professor with Inha University and a Research Scientist with the T-Brain, AI Center, SK Telecom, South Korea. His current research interests include multimodal learning, vision-language models, face understanding, and image segmentation.

IN KYU PARK (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering and computer science from Seoul National University, in 1995, 1997, and 2001, respectively. From September 2001 to March 2004, he was a member of Technical Staff with the Samsung Advanced Institute of Technology. Since March 2004, he has been with the School of Information and Communication Engineering, Inha University, where he is currently a Full Professor. From January 2007 to February 2008, he was an Exchange Scholar with Mitsubishi Electric Research Laboratories. From September 2014 to August 2015, he was a Visiting Associate Professor with the MIT Media Laboratory. From July 2018 to June 2019, he was a Visiting Scholar with the Center for Visual Computing, University of California at San Diego. His current research interests include computer vision and graphics, including 3D shape reconstruction from multiple views, image-based rendering, computational photography, deep learning, and GPGPU for image processing and computer vision. He is a member of ACM.