IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 November 2023, accepted 28 November 2023, date of publication 1 December 2023,
date of current version 12 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3338864

RESEARCH ARTICLE

Re-Aging GAN++: Temporally Consistent Transformation of Faces in Videos

FARKHOD MAKHMUDKHUJAEV[1], **(Member, IEEE), SUNGEUN HONG**[2], **(Member, IEEE), AND IN KYU PARK**[3], **(Senior Member, IEEE)**

[1]AI and Programming Department, Tashkent University of Information Technologies, Tashkent 100084, Uzbekistan
[2]Department of Immersive Media Engineering, Sungkyunkwan University, Seoul 03063, South Korea
[3]Department of Electrical and Computer Engineering, Inha University, Incheon 22212, 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 Korean Government through MSIT under Grant 2022R1A4A1033549, Grant NRF-2019R1A2C1006706, and Grant RS-2023-00211348; and in part by the Inha University Research Grant.

ABSTRACT The challenge of transforming the apparent age of human faces in videos has not been adequately addressed due to the complexities involved in preserving spatial and temporal consistency. This task is further complicated by the scarcity of video datasets featuring specific individuals across various age groups. To address these issues, we introduce Re-Aging GAN++ (RAGAN++), a unified framework designed to perform facial age transformation in videos utilizing an innovative GAN-based model trained on still image data. Initially, the modulation process acquires multi-scale personalized age features to depict the attributes of the target age group. Subsequently, the encoder applies Gaussian smoothing at each scale, ensuring a seamless frame-to-frame transition that accounts for inter-frame variations, such as facial motion within the camera's field of view. Remarkably, the proposed model demonstrates the ability to perform facial age transformation in videos despite being trained exclusively on image data. Our proposed method exhibits exceptional spatio-temporal consistency concerning facial identity, expression, and pose while maintaining natural variations across diverse age groups.

INDEX TERMS Video generation, age manipulation, GAN, spatio-temporal consistency.

I. INTRODUCTION

The process of changing the apparent age of a human face involves either making it look older or younger. This is done through age progression or regression, which uses a complex model trained on age distributions to manipulate a given face to match a target age. The target age acts as a conditional term that guides the model to produce facial images with characteristics appropriate for a certain age, as shown in Figure 1. However, video-based age transformation methods face challenges not encountered in image-based methods. These challenges arise due to the need to maintain consistency between frames and variations within frames to produce spatially and temporally consistent results. The output face must represent the same identity as the input face with age factors aligned according to the geometry of the face, including the pose and expression. Additionally, smooth transitions between adjacent frames are required to maintain temporal consistency.

Existing studies have mainly focused on transforming face age in images [1], [2], [3], [4], [5], [6], [7], [8]. However, the literature contains few discussions on transforming face age in video content, and only a limited number of approaches have been developed to handle video data, as mentioned in [9] and [10]. These methods can perform some modifications, including traversing a latent space, classification, and fine-tuning. Unfortunately, these methods often produce noticeable artifacts and unnatural or jittering effects, which can significantly reduce the overall quality and realism of the transformed videos. Additionally, the scarcity of video datasets containing faces of different age groups and specific subjects is a significant obstacle in the development of video-based facial age transformation methods. As we previously discussed, another important issue in this field is

The associate editor coordinating the review of this manuscript and approving it for publication was Yu-Huei Cheng.

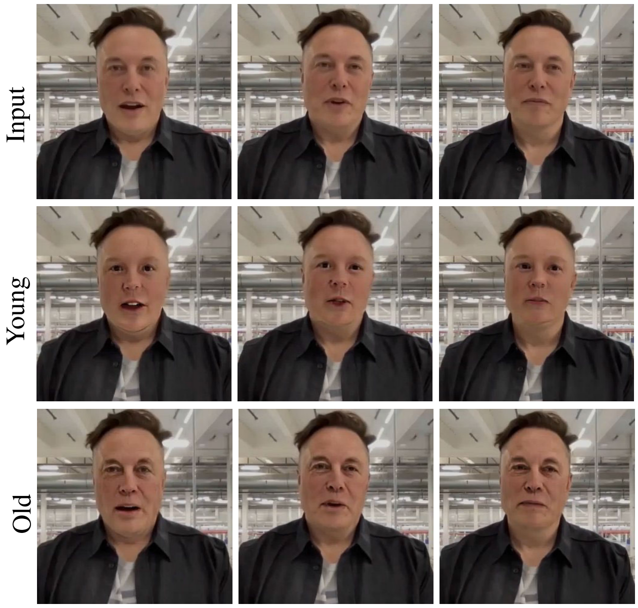

FIGURE 1. Age transformation results generated by the proposed method in video content. The first row shows the sample frames of the input video sequence, while the second and third rows show the result of age regression (*i.e.*, younger) and progression (*i.e.*, older), respectively.

the efficiency of deep learning based models, as mentioned in [11].

This study aims to propose a face age transformation method that can generate naturalistic faces appearing to be of a specific target age in video content, while handling intra- and inter-frame variations. To achieve this goal, we propose a new framework called RAGAN++, which considers video age transformation as an image translation problem. We were motivated by the success of an existing method with an age modulator module [7], and developed a more scalable framework to process video data without relying on external models. We maintain the self-guidance characteristics of the existing method using the age modulator, while also introducing an encoder-decoder structure with point modifications to handle variations within and between frames.

Our proposed framework consists of an encoder-decoder structure that integrates a multi-scale age modulator. This approach facilitates direct feature flow and avoids feature space alignment. To ensure visually realistic and accurate transformations, we designed the encoder to focus only on face regions in image frames, preserving necessary intra-frame variations like identity and background. We also added a Gaussian smoothing layer in each scale of the encoder to mimic inter-frame variations. Smoothness is considered as a perception of motion to simulate temporal consistency in videos from still images. Thus, our generator does not require labeled video sets and is applicable to real videos. By modulating identity features in a multi-scale manner, we apply specific age variations at each scale of image decoding using an age modulator. The main contributions of this work are summarized as follows.

- We propose a novel framework capable of transforming the apparent ages of faces in videos, even though it is trained only on static image data.
- We introduce an encoder to create smooth motion transitions between adjacent frames, which is critical for maintaining temporal consistency.
- Our multi-scale personalized age features enable age transformation while preserving the identity and appearance of the input face.
- Extensive experimental results show that our proposed method, which solely relies on image data, produces more realistic face age transformations in video content.

II. RELATED WORK

In this section, we introduce the existing face aging methods that are more relevant to our approach, while a comprehensive review of face aging works can be found in [12] and [13].

A. IMAGE-TO-IMAGE TRANSLATION

A wide variety of studies [3], [14], [15], [16], [17], [18], [19], [20], [21], [22] have extensively explored the use of GAN models to achieve greater realism in age transformation. However, these approaches often produce blurry results and distort the identity information of the input face, as pointed out by recent works [2], [23]. Consequently, methods that operate at higher resolutions have been proposed to address this issue. For instance, HRFAE [2] was developed to perform high-resolution age transformations using a simple procedure that re-weights the encoded features using the output of a single fully connected layer. Despois et al. [24] proposed a method that integrates aging maps into the decoding procedure through the adoption of SPADE blocks [25] to produce high-resolution facial images. However, these methods are mostly limited to age progression, and their ability to perform age regression is not always satisfactory.

There are several recent methods proposed for lifespan synthesis. LATS [1] is one such method that performs modulated convolutions on identity and injects a latent vector representing the target age, similar to StyleGAN2 [26]. However, LATS is limited to face synthesis only and does not incorporate background data, which can result in undesired artifacts. In contrast, the authors of [5] proposed a method that disentangles the characteristics of faces, such as shape, texture, and identity, to effectively preserve identity and model unique shapes and textures with respect to a target age. While this method provides better texture transformation and shape deformation, it suffers from the same issues as LATS. Another recent work [6] embeds an age estimator and personalized age embedding transformation modules into a regular encoder-decoder architecture to remove aging factors at the estimated age and obtain an embedding at the target age. However, this estimator may modify or shift identity traits in certain scenarios, limiting its effectiveness.

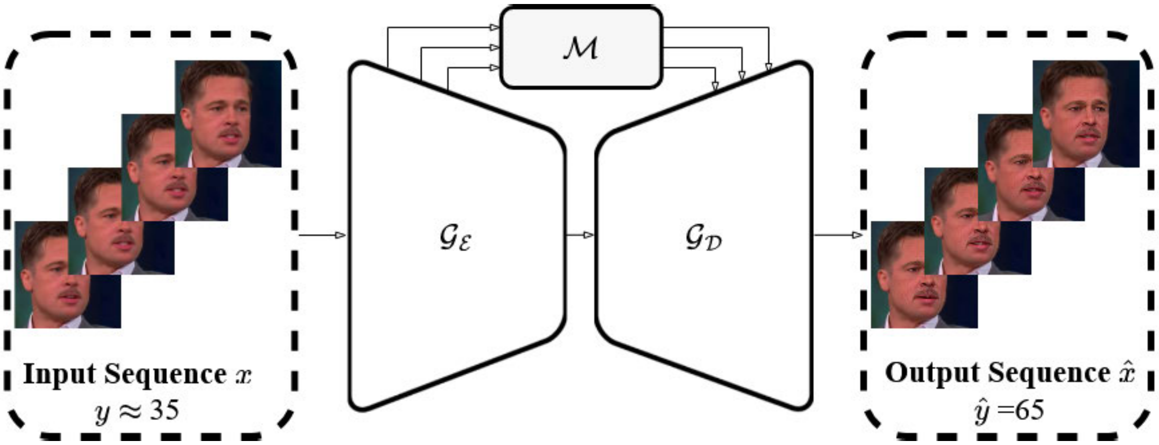
$\mathcal{M}$
$\mathcal{G}_{\mathcal{E}}$
$\mathcal{G}_{\mathcal{D}}$
Input Sequence x
$y \approx 35$
Output Sequence $\hat{x}$
$\hat{y}$ =65

FIGURE 2. Illustration of the proposed framework for face age transformation in videos. The proposed encoder and multi-scale modulations are designed to transform given input sequences to a target age $\hat{y}$ in a more visually realistic way.

B. VIDEO-TO-VIDEO TRANSLATION

StyleGAN is a highly advanced image generation method that has been proposed for generating highly realistic images [26], [27]. It has been used in various applications, including video-to-video translation, which is an extension of image-to-image translation for video content [9], [10]. To generate the intended content, several studies have employed pre-trained StyleGAN models while introducing diverse methods to find latent or manipulation directions of an image [28], [29]. However, finding and labeling these directions is a tedious process and is generally difficult to accomplish due to the nature of the StyleGAN space, particularly the intermediate latent $\mathcal{W}$ space. In light of this, Yao et al. [9] recently conducted a study to edit faces in real videos using StyleGAN. They inverted images to $\mathcal{W}^{+}$ space by using a dedicated encoder [29] to match latent spaces with a trained latent-code transformer. This allowed for more disentangled edits, including age, to be carried out.

Tzaban et al. [10] utilized StyleGAN to manipulate faces in video frames. They proposed a straightforward reconstruction-to-stitch pipeline, which incorporated both reconstruction and stitching to fine-tune the generator. The image inversion method [29] and finding image pivots using PTI [30] were combined in this method, enabling it to perform highly accurate yet highly editable reconstructions by simply editing the latent space in a given semantic direction. In contrast, an earlier work by Duong et al. [31] used deep reinforcement learning for face aging to perform video age progression, with the method performing manifold traversal from younger to older age regions given a latent representation of frames. Furthermore, Zoss et al. [32] proposed a practical, production-ready face re-aging network called FRAN that is controllable, temporally stable, and identity-preserving. However, this method only synthesizes the face region, with other parts such as hair and teeth remaining static regardless of aging. Furthermore, prior research in video-based age transformation has been limited due to the need for a significant amount of age-related face videos. This requirement makes it impractical to perform age transformation on a large scale. In contrast, a Gaussian smoothing layer was added to each level of the proposed encoder to replicate inter-frame variations and ensure temporal consistency, enabling the generator to be used with real videos without requiring labeled video datasets.

III. PROPOSED METHOD

A. OVERVIEW

In the context of image translation, the generator $\mathcal{G}$ is trained to convert input images x from one domain $\mathcal{A}$ to another domain $\mathcal{B}$, while the discriminator $\mathcal{D}$ learns to distinguish between real and fake images. Typically, the generator and discriminator are constructed using an encoder-decoder architecture, which up-samples and down-samples data to learn intricate image distributions. Our goal is to train a single generator $\mathcal{G}$ to produce an image $\hat{x}$ of a specific age $\hat{y} \in \mathcal{Y}$ from an input image $x \in \mathcal{X}$, where the generated image looks natural and maintains both intra- and inter-frame variations. To accomplish this, we propose an encoder that includes a Gaussian smoothing layer to smooth out motion transitions and we present multi-scale age modulations to facilitate naturalistic age manipulation. We provide an overview of our proposed framework in Figure 2. It is worth noting that our framework is trained on still images but can be applied to video content.

B. ENCODER

Recent studies have demonstrated that generators with an encoder structure are highly effective in producing naturalistic images. Most works incorporating down-sampling convolutional blocks designed as encoders to extract features from input data. An encoder $\mathcal{G}_{\mathcal{E}}$ is responsible for extracting the identity information $\mathcal{F}id$ of a given face x, as represented by the equation $\mathcal{F}id = \mathcal{G}_{\mathcal{E}}(x)$. This simple feature extraction supplies facial information at various levels. This approach is essential to maintain intra-frame variations of the target

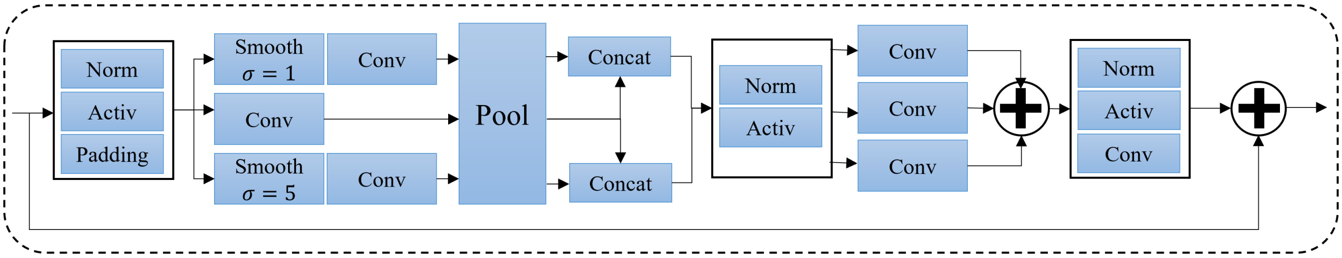

FIGURE 3. **Detailed architecture of the proposed encoder network.**

face, such as the person's identity, expression, and pose, to generate an image that appears to represent the same person as the output. Hence, we are interested in extracting only face-related features and avoid processing background variations that are unnecessary for face age transformation. To separate the former from the latter, we first crop the face region and then adopt a masking operation [7], [33], [34] at the initial image-to-feature layer of $\mathcal{G}_{\mathcal{E}}$ to avoid retaining information on the image background. This enables $\mathcal{G}$ to focus solely on the features of the target face, whereas the background can be reintegrated in the final image decoding layer.

However, a standard encoder designed for image translations may not suffice to handle interframe variations. As mentioned above, considering the motion of the input face between adjacent frames is important in manipulating its apparent age. Thus, we consider that the standard encoder part of $\mathcal{G}$ may tend to generate distinct faces where smooth motion transition is lost. This issue is discussed in greater detail in the description of the ablation studies carried out as part of this work. To address this, we modify $\mathcal{G}_{\mathcal{E}}$ according to the structure shown in Figure 3. Specifically, this structure adds Gaussian smoothing layers at each scale of the encoder alongside the standard feature-flow of the encoder. These multi-scale features provide a complementary representation encompassing the global and local characteristics of the face to help the encoder to mimic temporal variations in the generated images. We opt for Gaussian smoothing because Gaussian kernel filtering is widely used to remove undesirable fluctuations in video stabilization [35], [36]. However, in contrast to video stabilization, we aim to prevent $\mathcal{G}$ from producing abrupt changes of appearance in transformed adjacent frames by smoothing deep features of the encoder. Considering different types of fluctuations, we incorporate two Gaussian smoothing layers within each residual block. We empirically set σ_l and σ_h as low and high values, respectively, to mimic minimal and maximal smoothness (*i.e.*, slow and fast motion transitions in adjacent frames) on feature maps.

C. MULTI-SCALE MODULATIONS

Age modulations is an important aspect of this research. In contrast to existing methods, we utilize meaningful multi-scale features to facilitate better age transformation. Every scale of $\mathcal{G}_{\mathcal{E}}$ supplies coarse to-fine features that convey general information about the face, whereas existing methods [2], [7], [23] simply focus on bottleneck features. We consider that this approach leads to information loss, and address this issue by adopting a multi-scale approach in our age modulations. Specifically, our modulator takes identity features $\mathcal{F}_{id}$ from the encoder as usual, but produces its age-aligned counterpart by learning $\mathcal{F}_{age} = \mathcal{M}(\mathcal{F}_{id}, \hat{y})$ from each scale of $\mathcal{G}_{\mathcal{E}}$ and by considering the target age $\hat{y}$, where $\mathcal{F}_{age}$ is an N-dimensional vector that is further used in the decoder layers. To embed the target age into the modulations, we opt for the simplest option of establishing a bi-linear interaction between two trainable embedding functions E, E' and $\mathcal{F}_{id}$ formulated as $\mathcal{F}_{age} = \mathcal{F}_{id} + E(\hat{y}) + \mathcal{F}_{id} \odot E'(\hat{y})$. Thus, $\mathcal{G}$ learns the optimal age features for the input face at different scales, which enables the preservation of input details and the addition of personalized age characteristics. The proposed approach can perform joint training relatively easily because such modulations are incorporated into $\mathcal{G}$.

D. DECODER

We adopt a standard decoder architecture for $\mathcal{G}_{\mathcal{D}}$ to avoid any unnecessary complexity in learning to generate images. To produce age-transformed face outputs, the $\mathcal{G}_{\mathcal{D}}$ network operates on the identity $\mathcal{F}_{id}$ and multi-scale personalized age features $\mathcal{F}_{age}$ by performing $\hat{x} = \mathcal{G}_{\mathcal{D}}(\mathcal{F}_{id}, \mathcal{F}_{age})$. Through this operation, $\mathcal{G}_{\mathcal{D}}$ learns to preserve the identity of the input face while manipulating its age characteristics. Remarkably, the $\mathcal{F}_{age}$ feature addresses this requirement more effectively, because it belongs to the same identity and shares person-specific traits. Hence, the identity features $\mathcal{F}_{id}$ are self modulated by $\mathcal{F}_{age}$ through affine transformations that shift and center the features. We then reintegrate the background features into the image with the transformed facial features. To avoid any changes in both face and background information, we simply combine these features by addition prior to bringing the learned feature representation into the image domain. This enables the proposed approach to produce visually plausible image results. As a post-processing step, we adopt the well known Poisson blending [37] to combine the original input with the transformed face relatively easily.

E. DISCRIMINATOR

To lead $\mathcal{G}$ to learn meaningful information, we construct $\mathcal{D}$ following the successful architecture presented in [7] and [38]. To establish a fair competition between networks,

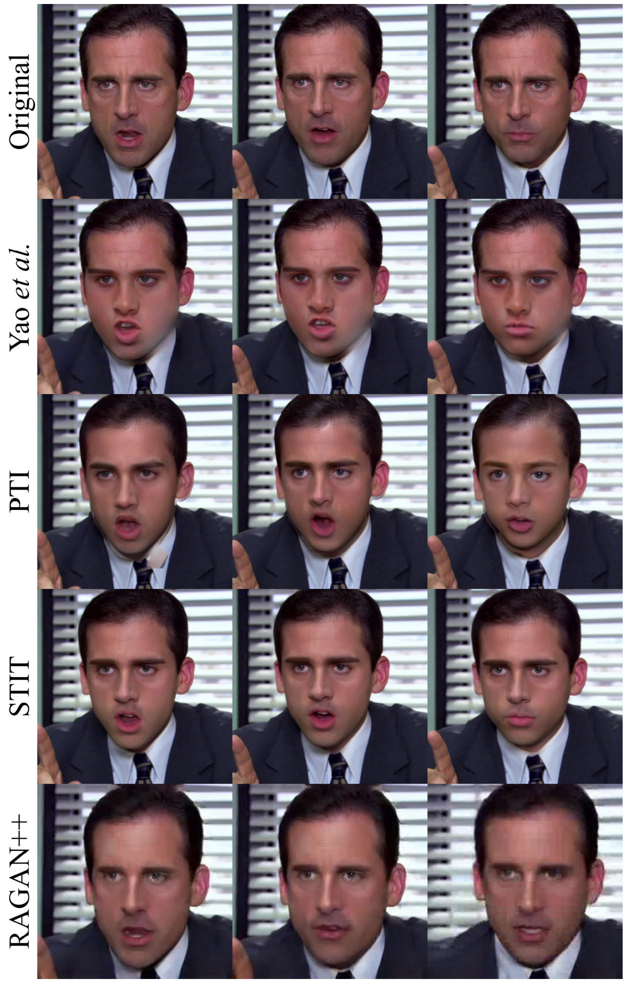

FIGURE 4. Visual comparison of the output of existing methods with that of our proposed method.

we train $\mathcal{D}$ on the image sets used for $\mathcal{G}$. Hence, the proposed model is an image-based discriminator that conducts a multi-task classification on the ages of faces in input images. Specifically, a classification layer consists of multiple output branches to differentiate among various ages. Each branch learns to distinguish whether an image is real or fake in its age domain by performing binary classification.

F. OBJECTIVE TERMS

Our framework aims to control the image generation process by transforming the inputs of the generator network through an age modulator. Because age labels are considered as a control factor, the framework must learn the proper condition-specific information of the target distribution. This requires the determination of objectives to appropriately supervise the framework. To this end, we utilize multiple objective functions, including adversarial, reconstruction, cycle consistency, and perceptual loss. To calculate the loss functions, we generate images based on an input image x and its real age label y, and a random target age $\hat{y}$. Accordingly, three different images are generated using the following:

$$\hat{x} = \mathcal{G}(x, \hat{y}), \quad x_{rec} = \mathcal{G}(x, y), \quad x_{cycle} = \mathcal{G}(\hat{x}, y), \qquad (1)$$

TABLE 1. Quantitative comparison of methods in terms of temporal consistency metrics proposed by [10].

Model	TL-ID↑	TG-ID↑
Latent Transformer [9]	0.976	0.811
PTI (optimization) [10]	0.933	0.901
STIT [10]	**0.996**	0.933
RAGAN++	0.957	**0.942**

where $\hat{x}$, x_{rec} and x_{cyc} are the transformed, reconstructed, and cycle-transformed images, respectively.

1) ADVERSARIAL LOSS

Our discriminator performs age-validity classification, the output of which is specific to the age domain. This approach can be considered as a class-conditioned $\mathcal{D}$. We define an adversarial loss as follows:

$$\mathcal{L}_{adv}(\mathcal{G}, \mathcal{D}) = \mathbb{E}_{x,y}\left[\log \mathcal{D}_y(x)\right] + \mathbb{E}_{x,y'}\left[\log\left(1 - \mathcal{D}_{y'}(x')\right)\right], \quad (2)$$

where $\mathcal{D}_y(\cdot)$ is the single output of $\mathcal{D}$ which belongs to y.

2) RECONSTRUCTION LOSS

During the training phase, the possibility that the target age $\hat{y}$ and the real age y of the input face x may fall into the same age group must be considered. In this case, $\mathcal{G}$ should produce a transformed image x $\hat{x}$ similar to the input x. To train the generator to handle such cases, we generate an image with its real age and use it to calculate a reconstruction loss as follows:

$$\mathcal{L}_{rec}(\mathcal{G}) = \|x - x_{rec}\|_1 \qquad (3)$$

3) CYCLE-CONSISTENCY LOSS

Shifts in apparent identity are an issue in face age transformation, and methods should address this in training $\mathcal{G}$. We consider that minimizing only (2) and (3) may not suffice. To address this circumstance, we adopt a cycle-consistency loss [39], [40], [41] in the training objectives. By doing so, the proximity of the age-transformed image $\hat{x}$ to the input x can be determined at inference. We lead the generator to focus on the difference between the age-transformed and input faces using the following expression:

$$\mathcal{L}_{cyc}(\mathcal{G}) = \|x - x_{cyc}\|_1 \qquad (4)$$

4) PERCEPTUAL LOSS

To encourage $\mathcal{G}$ to generate natural and perceptually realistic results, we adopt the learned perceptual image patch similarity LPIPS $(\mathcal{L}_{per})$ metric described in [42]. For this purpose, we utilize a VGG network pre-trained on the ImageNet dataset. We perform loss calculation by using the distance metric between the extracted features of the real and generated images.

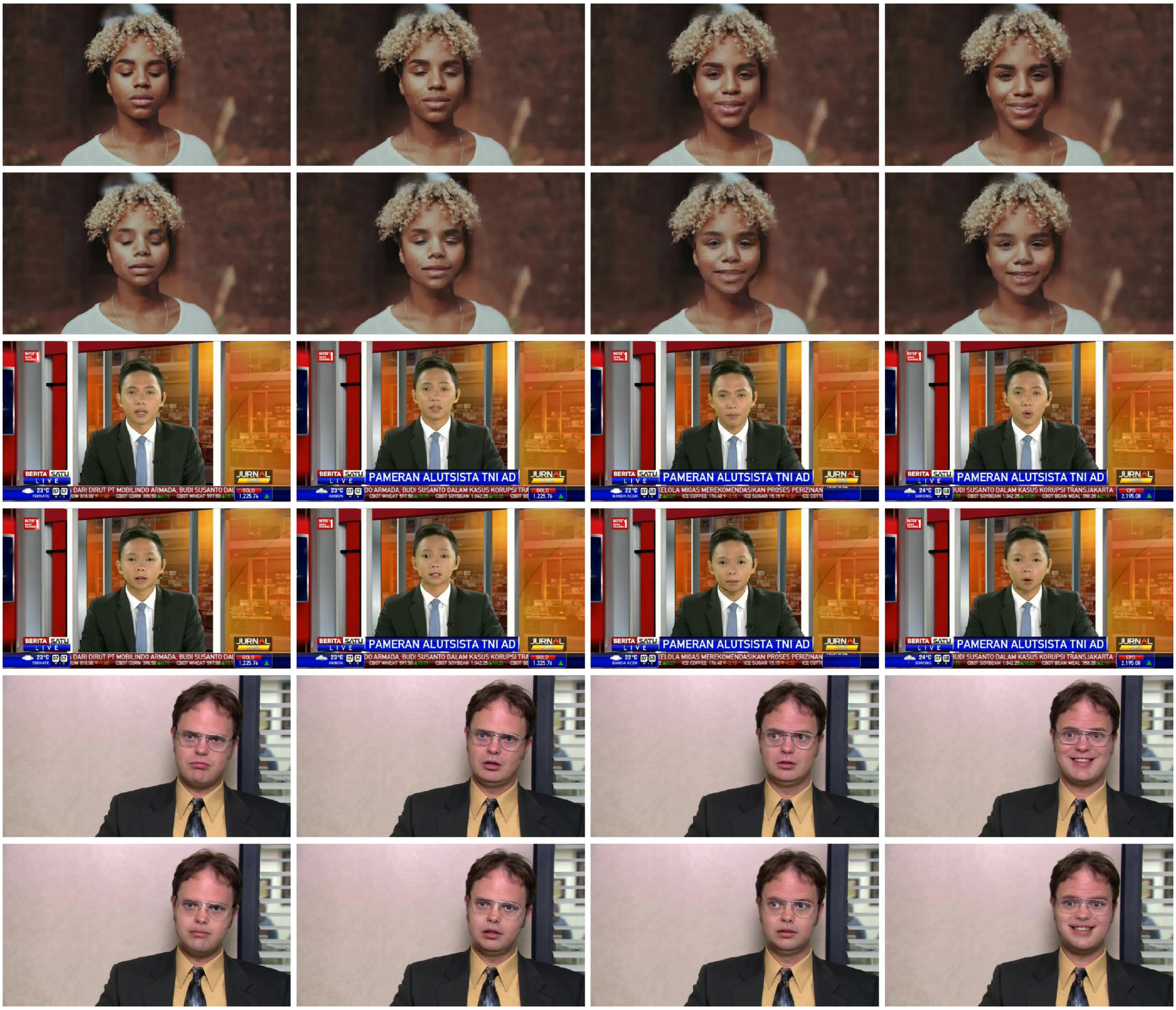

FIGURE 5. Qualitative age regression results produced by our method from in-the-wild videos. In each block, the first row shows the input sequence and the second shows the output sequences.

TABLE 2. Ablation study: Performance of our proposed methods when added to a baseline RAGAN model.

Method	FID↓	FVD↓
Baseline RAGAN	33.1	78.8
+ Temporal encoder	27.4	70.5
+ Multi-scale modulations	**26.7**	**68.2**

5) FULL OBJECTIVE

The following expression defines the overall objective considered to optimize $\mathcal{G}$ and $\mathcal{D}$:

$$\min_{G}\max_{D} \lambda_{adv}\mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cyc}\mathcal{L}_{cyc} + \lambda_{per}\mathcal{L}_{per} \quad (5)$$

where λ_{adv}, λ_{rec},, λ_{cyc} and λ_{per} are the weights necessary to avoid side effects of different losses in the training phase.

IV. EXPERIMENTAL SETUP

A. DATASET

We train the framework on the FFHQ dataset [27], which is labeled for the age transformation task. This dataset comprises images from 70,000 different identities in 10 age groups. By following the strategy presented in [1], we prepare training and testing datasets that omit images with low confidence scores for the included labels. The resolution of all the images utilized in the training phase and evaluations is set to 256 × 256 pixels as a standard resolution used in existing studies.

B. HYPERPARAMETERS

The proposed framework is trained on a single NVIDIA RTX A6000 GPU (48 GB) with a batch size of 12 for 30 epochs.

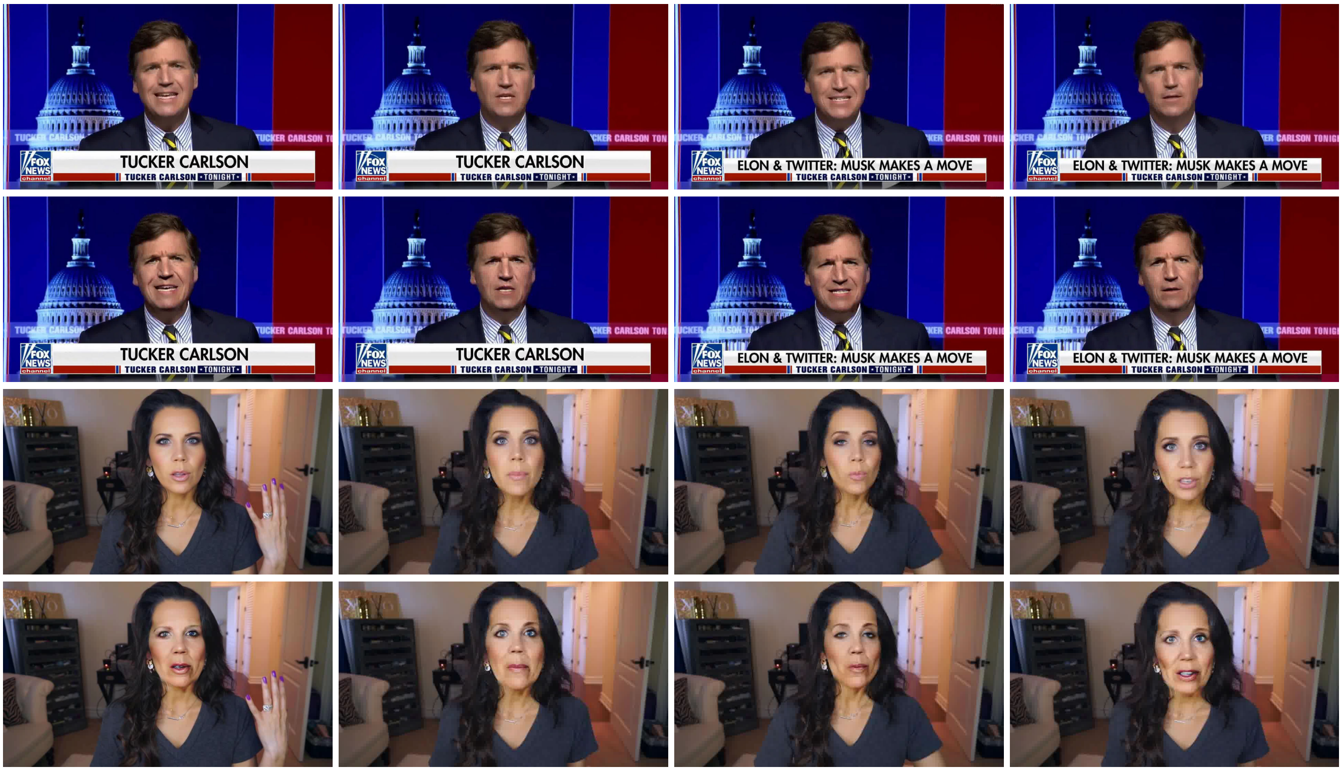

FIGURE 6. Qualitative age progression results produced by our method from in-the-wild videos. In each block, the first row shows the input sequence and the second shows the output sequences.

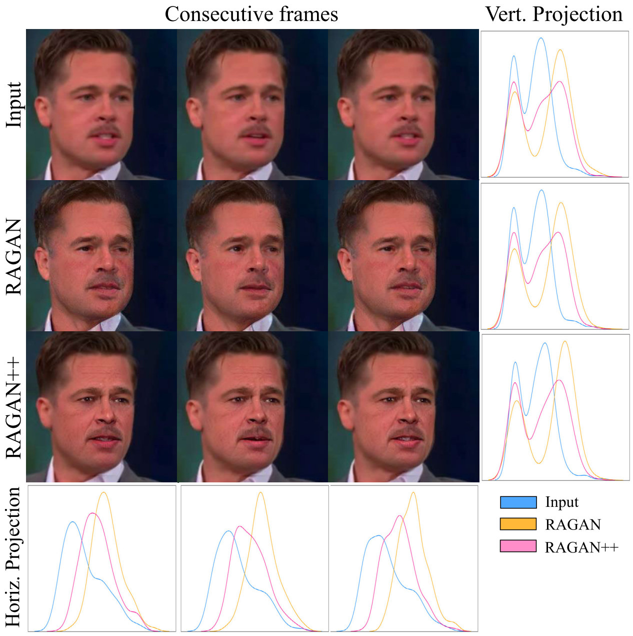

FIGURE 7. Age transformation results on consecutive frames generated by RAGAN and RAGAN++.

We utilize the well-known Adam [43] optimizer with the following parameter settings: $\beta_1 = 0.0$, $\beta_2 = 0.99$, and $\eta = 10^{-4}$. We also incorporate R1 regularization [44] in the training phase. In addition, the learning rate scheduler is used for the generator and discriminator. We train the model with $\lambda_{rec} = 10$, $\lambda_{cyc} = 1$, and $\lambda_{adv} = 1$ during the initial 2 epochs. Then, we switch $\lambda_{rec} = 1$ to help the generator focus on image reconstruction.

In the quantitative evaluations, we evaluate the generated images in terms of identity preservation and age modification. For this purpose, we utilize the Frechét inception distance (FID) [45] and Frechét video distance (FVD) [46] metrics, which evaluate the discrepancy between distributions in the image and video domains, respectively. In addition, we also adopt the identity preservation metrics proposed in [10]. To demonstrate the effectiveness of our method, we apply it to a range of in the wild videos gathered from popular publicly available content. These include challenging scenes characterized by complex backgrounds and considerable movement.

V. EXPERIMENTAL RESULTS

A. QUALITATIVE EVALUATION

We perform a qualitative comparison between our proposed approach and existing methods for video data. Specifically, we compare our results with those of Yao et al. [9], PTI [30], and STIT [30]. A qualitative comparison is presented in Figure 4. Our framework is able to generate visually plausible output sequences while modifying the ages of the input faces realistically. We consider that the results of or approach look better than or as good as those of recent state-of-the-art methods in terms of their subjective appearance. Overall, our simple yet effective single framework is able to maintain

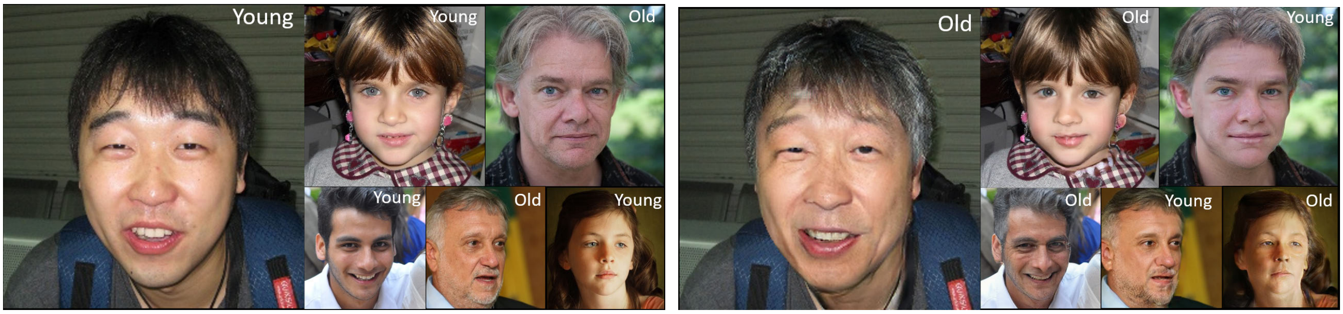

FIGURE 8. Image-to-image translation results using RAGAN++.

intra- and inter-frame variations while transforming faces according to target ages.

We apply the proposed method to real-world videos. Figure 5 and Figure 6 show example of the outputs, which is evaluated for age regression and progression cases, respectively. The readers also might refer to the supplementary material for the video results. As observed from the comparison results, our proposed method succeed in adding and removing aging-related lines, such as wrinkles and facial hair, without affecting the consistency between adjacent frames. However, it should be noted that we also observed a certain limitation of the proposed method in handling extreme poses, *i.e.*, faces that appear in images from other than a front-facing perspective, which is due to the limitation of the model having been trained on a dataset that mostly includes images in which the input person is roughly facing the camera.

B. QUANTITATIVE EVALUATION

For a quantitative evaluation, we follow the existing work of [10], which proposed a method to evaluate the coherence of generated videos. Specifically, we evaluate the performance in terms of two metrics that consider temporally local (TL-ID) and temporally global (TG-ID) identity preservation. The former (TL-ID) assesses the local consistency of a video by comparing the properties of pairs of adjacent video frames using an identity-detection network [47]. A method with higher TL-ID scores is more likely to yield results that are smooth and free of significant local identity jitters or artifacts. The latter (TG-ID) uses the same identity detection network as well as an averaging strategy to compare all possible pairs (regardless of adjacency) of video frames to determine their similarity. This metric attempts to detect slow but consistent identity drift and capture long-range coherence. For both measures, a score of 1 would mean that the method effectively preserves the identity of the source video consistently.

Table 1 presents a comparison of the results of our proposed framework with those of Yao et al. [9], PTI [30], and STIT [10]. It is observed that our method achieves better results in terms of both metrics compared to the others, which suggests its temporal coherence. Of note, these results also show the higher quality of the generated videos. However, the performance of our method is slightly lower than that of STIT in terms of TL-ID. We attribute this result to our generator learning to introduce more local-level age transformation on the face regions.

C. ABLATION STUDY

We begin our ablation studies by demonstrating the efficiency of our proposed encoder in comparison to the existing RAGAN [7]. We transform three consecutive frames to an older age (*i.e.*, $\hat{y} = 55$) using the existing and our proposed methods. Figure 7 shows the input along with the transformed frames. Although RAGAN generates an image of a person according to the target age, face regions showing motion are damaged or lost, specifically the mouth region. We also calculate the vertical and horizontal projections of these images to show how the input and transformed distributions diverged. It is observed that the distributions of the images output by RAGAN++ in both directions are close to those of the input images, which suggests a need for smoother transitions.

We conducted an ablation study to demonstrate the effectiveness of each component of our proposed pipeline. The purpose of this study was to validate the contribution of different components by switching and progressively adding them to the framework. To this end, we trained a model with the baseline encoder and gradually added the proposed components. Table 2 presents the effects of replacing the baseline encoder with the proposed encoder and generating age-transformed images by adding multi-scale modulations only. The results show that the proposed encoder yielded a significant improvement compared to the baseline encoder. This improvement can be attributed to the better representation of texture transformations such as wrinkles, which are characteristic of older faces, due to the smoothing process added in the temporal encoder. On the other hand, multi-scale modulations resulted in improved age transformations of the faces. Overall, these findings support the effectiveness of our proposed pipeline and its components for generating visually realistic age transformations in video content.

D. IMAGE-TO-IMAGE TRANSLATIONS

We conducted an experiment to show the effectiveness of our proposed framework for image-to-image translation by performing a simple face age transfer. Our aim was to demonstrate how our approach could be applied in various image translation tasks. To evaluate the quality of our approach, we compared our results with those obtained using the baseline RAGAN model. The qualitative results, which are presented in Figure 8, indicate that our proposed approach did not result in any significant reduction in the quality of the generated images. Furthermore, we observed an improvement in the quality of the mapping of one expression to another when using our proposed model. This suggests that our approach could be applied in other image translation tasks where the quality of the mapping is an important factor. Overall, the results of our experiment provide evidence that our proposed framework is a promising approach for image-to-image translation tasks.

VI. CONCLUSION

In this study, we presented a novel framework for age transformation called RAGAN++. Our proposed method enables the generation of visually consistent aging effects on faces in video content while preserving the intra- and inter-frame variations present in the videos. By utilizing a modulation process, we were able to spatially transform the age of the input face and reliably introduce the characteristics of the target age in a target person's appearance. We also incorporated Gaussian smoothing layers into the encoder structure to simulate smooth facial motion transitions in a more efficient manner, resulting in temporally consistent results. Despite being trained on image data, our single framework was able to effectively transform face age on video content, demonstrating its applicability to real-world scenarios. Moreover, the visual perception of the transformed image results of our proposed method was superior to that of existing methods.

REFERENCES

[1] R. Or-El, S. Sengupta, O. Fried, E. Shechtman, and I. Kemelmacher-Shlizerman, "Lifespan age transformation synthesis," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 739-755.

[2] X. Yao, G. Puy, A. Newson, Y. Gousseau, and P. Hellier, "High resolution face age editing," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 8624–8631.

[3] Q. Li, Y. Liu, and Z. Sun, "Age progression and regression with spatial attention modules," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11378–11385.

[4] D. Yadav, N. Kohli, M. Vatsa, R. Singh, and A. Noore, "Age gap reducer-GAN for recognizing age-separated faces," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10090–10097.

[5] S. He, W. Liao, M. Y. Yang, Y.-Z. Song, B. Rosenhahn, and T. Xiang, "Disentangled lifespan face synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3857–3866.

[6] Z. Li, R. Jiang, and P. Aarabi, "Continuous face aging via self-estimated residual age embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15003–15012.

[7] F. Makhmudkhujaev, S. Hong, and I. Kyu Park, "Re-aging GAN: Toward personalized face age transformation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3888–3897.

[8] 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, doi: 10.1145/3450626.3459805.

[9] X. Yao, A. Newson, Y. Gousseau, and P. Hellier, "A latent transformer for disentangled face editing in images and videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 13769–13778.

[10] R. Tzaban, R. Mokady, R. Gal, A. H. Bermano, and D. Cohen-Or, "Stitch it in time: GAN-based facial editing of real videos," 2022, *arXiv:2201.08361*.

[11] S. A. Ajagbe, O. A. Oki, M. A. Oladipupo, and A. Nwanakwaugwu, "Investigating the efficiency of deep learning models in bioinspired object detection," in *Proc. Int. Conf. Electr., Comput. Energy Technol. (ICECET)*, Jul. 2022, pp. 1–6.

[12] M. Grimmer, R. Ramachandra, and C. Busch, "Deep face age progression: A survey," *IEEE Access*, vol. 9, pp. 83376–83393, 2021.

[13] H. Pranoto, Y. Heryadi, H. L. H. S. Warnars, and W. Budiharto, "Recent generative adversarial approach in face aging and dataset review," *IEEE Access*, vol. 10, pp. 28693–28716, 2022.

[14] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face aging with conditional generative adversarial networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2089–2093.

[15] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4352–4360.

[16] X. Tang, Z. Wang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7939–7947.

[17] J. Song, J. Zhang, L. Gao, X. Liu, and H. T. Shen, "Dual conditional GANs for face aging and rejuvenation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 899–905.

[18] H. Yang, D. Huang, Y. Wang, and A. K. Jain, "Learning face age progression: A pyramid architecture of GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 31–39.

[19] P. Li, Y. Hu, Q. Li, R. He, and Z. Sun, "Global and local consistent age generative adversarial networks," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1073–1078.

[20] P. Li, Y. Hu, R. He, and Z. Sun, "Global and local consistent wavelet-domain age synthesis," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 11, pp. 2943–2957, Nov. 2019.

[21] Y. Liu, Q. Li, Z. Sun, and T. Tan, "A3GAN: An attribute-aware attentive generative adversarial network for face aging," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2776–2790, 2021.

[22] Y. Liu, Q. Li, and Z. Sun, "Attribute-aware face aging with wavelet-based generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 11869–11878.

[23] Z. He, M. Kan, S. Shan, and X. Chen, "S2GAN: Share aging factors across ages and share aging trends among individuals," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9440–9449.

[24] J. Despois, F. Flament, and M. Perrot, "AgingMapGAN (AMGAN): High-resolution controllable face aging with spatially-aware conditional GANs," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 613–628.

[25] 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. 2332–2341.

[26] 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.

[27] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[28] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "GANSpace: Discovering interpretable GAN controls," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9841–9850.

[29] E. Richardson, Y. Alaluf, O. Patashnik, Y. Nitzan, Y. Azar, S. Shapiro, and D. Cohen-Or, "Encoding in style: A styleGAN encoder for image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2287–2296.

[30] D. Roich, R. Mokady, A. H. Bermano, and D. Cohen-Or, "Pivotal tuning for latent-based editing of real images," 2021, *arXiv:2106.05744*.

[31] C. N. Duong, K. Luu, K. G. Quach, N. Nguyen, E. Patterson, T. D. Bui, and N. Le, "Automatic face aging in videos via deep reinforcement learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10005–10014.

[32] G. Zoss, P. Chandran, E. Sifakis, M. Gross, P. Gotardo, and D. Bradley, "Production-ready face re-aging for visual effects," *ACM Trans. Graph.*, vol. 41, no. 6, pp. 1–12, Dec. 2022.
[33] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[34] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "MaskGAN: Towards diverse and interactive facial image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5548–5557.
[35] M. Roberto e Souza, H. D. A. Maia, and H. Pedrini, "Survey on digital video stabilization: Concepts, methods, and challenges," *ACM Comput. Surv.*, vol. 55, no. 3, pp. 1–37, Mar. 2023.
[36] P. Rawat and M. D. Sawale, "Gaussian kernel filtering for video stabilization," in *Proc. Int. Conf. Recent Innov. Signal Process. Embedded Syst. (RISE)*, Oct. 2017, pp. 142–147.
[37] P. Pérez, M. Gangnet, and A. Blake, "Poisson image editing," in *Proc. ACM SIGGRAPH Papers*, Jul. 2003, pp. 313–318.
[38] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha, "StarGAN v2: Diverse image synthesis for multiple domains," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8185–8194.
[39] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1–9.
[40] 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. (ICCV)*, Oct. 2017, pp. 2242–2251.
[41] 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.
[42] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[44] L. Mescheder, S. Nowozin, and A. Geiger, "Which training methods for GANs do actually converge?" in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–10.
[45] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
[46] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "FVD: A new metric for video generation," in *Proc. ICLR*, 2019, pp. 1–9.
[47] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

FARKHOD MAKHMUDKHUJAEV (Member, IEEE) received the B.S. and M.S. degrees from the Tashkent University of Information Technologies (TUIT), Uzbekistan, in 2012 and 2014, respectively, and the Ph.D. degree in computer science and engineering from Kyung Hee University, South Korea, in 2019. From October 2019 to September 2022, he was a Postdoctoral Researcher with the Artificial Intelligence Convergence Research Center, Inha University. He is currently an Assistant Professor at TUIT. His current research interests include image synthesis using generative adversarial networks and facial attribute analysis and recognition.

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 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 (SNU), in 1995, 1997, and 2001, respectively. From September 2001 to March 2004, he was a Member of the Technical Staff with the Samsung Advanced Institute of Technology (SAIT). Since March 2004, he has been with the Department 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 the Mitsubishi Electric Research Laboratories (MERL). 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.