Adding Aging While Maintaining Facial Animation

Brendon Fang (blfang) & Max Fang (msfang) & Rosemary Yang (rey)
10-423/623 Generative AI Course Project

December 13, 2024

1 Introduction

This project aims to extend the GANimation framework by integrating age progression and regression into its anatomically aware facial animation model. Motivated by the need for realistic, age-conditional facial expression synthesis, our approach introduces age-related parameters alongside existing Action Unit (AU) conditioning. This extension enables the generation of photorealistic images that combine facial aging transformations with dynamic emotional expressions, all while preserving identity and key facial features. We used high-quality data sets, such as FFHQ and CACD, and evaluated our model qualitatively.

Our proposed method builds upon the GANimation framework, which uses AU intensities to generate realistic facial expressions. By modifying the conditioning vector to include age parameters, we seek to create a seamless integration of age transformations without compromising the accuracy of facial expressions. The expected result is a model capable of generating age-progressed and age-regressed faces that retain the subject's identity, while also manipulating facial expressions with high realism and anatomical coherence. This work represents a novel application of GAN-based techniques for joint age and emotion manipulation in facial animation.

2 Dataset/Task

The proposed work will utilize the following datasets, all of which contain high-quality headshots of people, suitable for facial animation and transformation tasks:

FFHQ (Flickr-Faces-HQ) Dataset: This dataset provides high-resolution images of faces, featuring diverse age groups, ethnicities, and styles. It can be accessed from `https://paperswithcode.com/dataset/ffhq`[ffh].

CACD (Cross-Age Celebrity Dataset): This dataset includes celebrity headshots captured across different age ranges, making it particularly suitable for age-related transformations. The data set can be accessed from `https://paperswithcode.com/dataset/cacd`[cac].

Our objective is to enhance the capabilities of the GANimation model by introducing an aging dimension to its existing anatomically-aware facial animation framework. GANimation is designed for facial expression manipulation and was presented in the paper "GANimation: Anatomically-aware Facial Animation from a Single Image".

The existing codebase for GANimation is publicly available at `https://github.com/albertpumarola/GANimation` [Pumarola, 2018], and our work will build upon this implementation to incorporate age progression and regression features. An evaluation metric for this project will involve a qualitative assessment of the generated images. The synthesized outputs will be compared to real images from the FFHQ, and CACD datasets to verify their realism and adherence to age-related transformations. This comparison will focus on visual consistency, anatomical plausibility, and the preservation of key facial attributes across aging transformations.

3 Related Work

Generative Adversarial Networks (GANs) are a class of generative models that consist of two neural networks: a generator G and a discriminator D, trained in a min-max framework. The generator synthesizes realistic data, while the discriminator distinguishes between real and generated samples. Numerous studies have extended and tailored GAN architectures for specific applications, showcasing their adaptability across different modalities and domains.

Karras [2018] introduced a training methodology for GANs that progressively grows the generator and discriminator networks. This process begins with low-resolution layers and incrementally adds higher-resolution layers, enabling the model to capture large-scale structures before focusing on finer details. The approach improves training stability and image quality. Additionally, the authors proposed a new evaluation metric that balances image quality and diversity, providing a more holistic assessment of GAN performance.

Deng J. [2022] proposed ArcFace, a loss function designed to enhance the discriminative capabilities of deep face recognition by incorporating an additive angular margin. This in-

novation improves inter-class separability, leading to state-of-the-art performance in face recognition tasks. Beyond recognition, the pre-trained ArcFace model demonstrates generative capabilities by synthesizing identity-preserving face images without additional generators or discriminators. It leverages network gradients and Batch Normalization priors to synthesize faces for unseen identities, underscoring its versatility in discriminative feature embedding and face generation.

Li [2020] introduced FaceShifter, a two-stage framework designed for high-fidelity and occlusion-aware face swapping. The Adaptive Embedding Integration Network (AEI-Net) generates identity-preserving outputs by integrating multi-level attributes from the target image, while the Heuristic Error Acknowledging Network (HEAR-Net) refines occlusions to enhance the final result. This method excels in maintaining the subject's identity while adapting attributes such as pose, expression, and lighting.

Shen [2020] proposed InterFaceGAN, a framework for semantic face editing through latent space interpretation. This method identifies semantic attributes in the latent space of GANs, enabling precise control over facial attribute editing even with fixed, pre-trained models. InterFaceGAN effectively transforms unconditional GANs into controllable models, allowing fine-grained manipulation of attributes such as age, gender, and expression.

Duarte [2019] in Wav2Pix: Speech-Conditioned Face Generation Using Generative Adversarial Networks, explored generating face images conditioned on raw speech inputs. This cross-modal task required the integration of a speech encoder into both the generator and discriminator, enabling the model to synthesize faces based on audio features. The approach highlights the potential of GANs in bridging visual and auditory domains for multimodal generation.

Li [2021] in Anime Face Generation with GANs, addressed the challenge of translating human portraits into anime-style images. Their model employs a generator that takes both a source image (a human portrait) and a reference image (a target anime style) to synthesize a new image combining features from both inputs. The discriminator uses two branches: one to ensure consistency with the reference image and another to assess realism relative to the source image. Di [2018] in GP-GAN: Gender-Preserving GAN for Synthesizing Faces from Landmarks, proposed a GAN-based approach to generate gender-preserving face images from facial landmarks. Their model captures the joint distribution of facial landmarks and corresponding face images, ensuring that the synthesized faces retain gender-specific features inherent to the input.

Yao X. [2020] in High-Resolution Face Age Editing, tackled the challenge of facial age transformation while maintaining high-quality ./outputs. Their architecture, based on an auto-encoder framework, incorporates an encoder, a feature modulation block, and a decoder. The key innovation lies in the age transformer, which modifies input faces to reflect a target age while preserving high-resolution details. This approach enables robust and realistic age transformations suitable for practical applications.

4 Methods

The baseline GANimation model consists of a generator that accepts an input image and target AU conditions to output a synthesized image and an attention mask. These outputs ensure that AU transformations are applied selectively to the appropriate regions of the face. Key loss functions, including cycle-consistency loss, mask smoothness loss, and AU consistency loss, help guide the generator to align its outputs with target AU vectors while preserving irrelevant facial features. The discriminator in the baseline GANimation model distinguishes real images from generated ones and conditions on AU vectors to enforce semantic alignment between input conditions and the generated outputs. While effective for AU-driven facial transformations, this architecture does not inherently support multi-attribute editing, such as age, gender, or accessory-based transformations.

To enable age-based editing, we introduced a modulation network that processes age inputs into a style embedding of 128 dimensions. This embedding is combined with the AU vector to create a unified condition vector that drives the generator. The modulation network ensures that age features are processed separately from AU conditions, minimizing entanglement between transformations. The unified condition vector is then spatially expanded and concatenated with the input image, allowing the generator to apply pixel-wise transformations informed by both AUs and age. This approach ensures precise control over age-related modifications, such as wrinkles or changes in skin tone.

To improve the discriminator's ability to capture fine-grained details, we integrated DisPatchGAN, which evaluates the realism of smaller image patches rather than the entire image. PatchGAN focuses on local features, such as texture and wrinkles, critical for capturing age-specific details. This modification simplifies the discriminator's task and stabilizes adversarial training. We also introduced a toggle mechanism to enable or disable the use of DisPatchGAN, so that we could focus on specific local areas during experimentation.

In addition to DisPatchGAN, we incorporated a VGG-based perceptual loss to preserve high-level semantic information. By comparing feature representations of real and generated images in a pre-trained VGG network, the perceptual loss ensures that transformations maintain the identity

and overall structure of the face. This loss adds onto pixel-level losses and adversarial feedback, resulting in more realistic and coherent outputs.

Finally, the discriminator was adapted to handle multiple attributes simultaneously, including realism, AU alignment, and age consistency. AU alignment was enforced through an AU consistency loss, implemented as a mean squared error, which ensures that the AUs predicted by the discriminator for the fake image match the desired AU conditions. Similarly, age consistency was achieved by introducing age-specific outputs in the discriminator, with age loss defined as mean squared error for the numerical age annotation.

5 Experiments

Figure 1: Baseline GANimation Model Results

The core objective of our experiments was to evaluate the GANimation framework's ability to incorporate age modulation and produce visually coherent results while maintaining Action Unit (AU) alignment and realism. We conducted experiments on a subset of 1,000 images from a larger dataset containing 70,000 samples. This decision to use a reduced dataset was driven by the need for rapid prototyping and resource constraints, allowing us to validate our model's functionality and optimize hyperparameters efficiently before scaling to the full dataset. The pilot study provided a manageable framework for debugging and qualitative assessments without incurring excessive computational costs.

Our experiments were conducted on an AWS EC2 instance using GPU. This configuration allowed us to process high-resolution images efficiently and handle the computational demands of training and validating the model.

Figure 2 demonstrates isolated age progression and regression without AU activations, as implemented in the extended GANimation framework.

The age-extended GANimation model was trained for 20

Figure 2: Isolated Age Progression and Regression without AU activations as implemented in the extended GANimation framework

Figure 3: Emotion and Age Manipulation on the extended GANimation framework, across a range of ages and AU intensities

epochs in two stages. In the first stage, we conducted 10 epochs on the FFHQ dataset with a batch size of 8 and a learning rate $\alpha = 1 \times 10^{-4}$ In the second stage, we refined the model with 15 additional epochs, selecting only 1024×1024 images from the dataset, with a reduced batch size of 2 and a smaller learning rate $\alpha = 1 \times 10^{-5}$ to stabilize training. Training required approximately 2.5 hours per epoch for the first stage and around 3.5 hours per epoch for the second stage. Validation runs were significantly faster, requiring approximately 30 minutes per dataset pass.

Figure 3 illustrates the outputs of our model as we combine AU intensities with different age transformations. The age modulation network successfully applies age-specific transformations (e.g., adding wrinkles and skin tone changes) while maintaining the underlying expressions dictated by the AU conditions. The results demonstrate consistent progression across age levels (e.g., from 25 to 69 years) while preserving facial features and spatial coherence. However, the generated images exhibited several shortcomings, including blurriness, uniform application of wrinkle patterns, and a lack of localized transformations.

These issues can largely be attributed to the limitations of the pilot experiment. The model was trained on a small subset of 1,000 images from a larger dataset of 70,000 images. While this subset allowed for quicker experimentation, the

reduced diversity in the training data limited the model's ability to generalize age-specific features and transformations. Furthermore, the age modulation network processes age as a single-dimensional vector transformed into a style embedding, which oversimplifies the complex relationship between age and facial features. This approach leads to generalized transformations rather than region-specific aging effects.

Another contributing factor is the design of the loss functions. While mean squared error (MSE) and cross-entropy were effective for aligning the output with target attributes, they did not explicitly guide the model to localize aging transformations. Additionally, the attention mechanism in the generator, though useful for blending transformations with input images, was not sufficiently refined to prioritize age-relevant areas of the face, further limiting the spatial specificity of the transformations. The challenges inherent in GAN training, particularly the difficulty in generating high-frequency details, also played a role in the blurriness of the outputs. The model prioritized global realism over fine-grained features to satisfy the discriminator, which may have occurred because our GAN was trained on a small dataset.

6 Research Log

Initially, the focus was on adding age as an additional feature to the original GANimation model. However, we wanted to expand to include three attributes: age, gender, and accessories, with each team member taking responsibility for implementing one feature. This decision required extending the model's architecture to support multi-attribute transformations while preserving its original capability to manipulate expressions through action units (AUs).

In addition to issues with age annotations, we faced technical difficulties while setting up OpenCV to handle Action Unit (AU) processing. OpenCV is a package that helps extract AU features, and has a lot of issues with installation. Many of the issues we encountered during the installation were documented online, but none were resolved, even though other researchers had experienced similar problems. After several attempts and workarounds that set us behind, we managed to get OpenCV running but were forced to pivot our dataset strategy to one that worked better with this pipeline.

Specifically for age, we implemented a dedicated modulation network designed to process the age feature independently. The idea was to allow the generator to handle age transformations with its own set of weights, minimizing entanglement with other attributes and maintaining flexibility for future expansions. This architecture provided the model with the ability to generate age-specific transformations such as adding wrinkles or altering skin tone.

However, the initial implementation of the age modulation network presented challenges. While it could independently apply some aging effects, the results lacked targeted transformations and realism. For example, wrinkles appeared uniform across the face instead of being localized to areas such as the eyes, mouth, or forehead, and there was no greying hair or skin discoloration. To improve these results, we incorporated Dis_PatchGAN and a VGG-based perceptual loss into the model. The Dis_PatchGAN discriminator enhanced the spatial precision of the transformations by focusing on localized patches of the face, while the VGG loss encouraged the preservation of high-level facial features, improving the realism of the generated images. These changes significantly improved the model's ability to generate smooth and realistic age transformations, as demonstrated in Figure 4.

Despite these improvements, combining age and AU (action unit) transformations presented another layer of challenges. While the Dis_PatchGAN discriminator effectively trained the age modulation network, it didn't fully integrate with the original GANimation discriminator. This created some inconsistencies in the combined results, as seen in Figure 5. The combined outputs did not have the detail or the localization that the isolated age model had.

In this respect, we also considered the possibility of directly manipulating AUs within the model to simplify the architecture. In hindsight, directly manipulating AUs within the model might have been a more straightforward approach. However, this would have required significant architectural changes to the generator, potentially entangling AUs with other features in unintended ways. By keeping attribute-specific transformations independent, the model can be easily adapted to include additional features in the future without significant architectural changes.

In trying to implement gender within the model, we encountered difficulties integrating a conditional vector. The process of encoding gender effectively proved more complex than anticipated, requiring further exploration of suitable representation methods. Additionally, implementing the gender-specific action unit architecture presented another layer of complexity. The binary nature of gender, combined with its intricate interplay with facial features, made disentangling these elements particularly difficult. This highlighted the need for a more nuanced approach to effectively incorporate gender without compromising the architecture's overall functionality.

An attempt was also made to extend the GANimation framework to incorporate accessory transformations, such as glasses and hats, alongside facial expression manipulation driven by Action Units (AUs). While initial efforts focused on adapting the AU-based architecture to include accessory attributes, several challenges ultimately hindered the

successful implementation of accessory generation. AUs, being anatomically defined, are inherently limited to modeling facial muscle movements, which made it difficult to encode non-anatomical features such as accessories in the same framework. Attempts to integrate accessory attributes into the conditional input vector resulted in entanglement with facial features, leading to inconsistent and anatomically implausible outputs.

Efforts to adapt the attention mechanism for accessory-specific regions, such as the forehead for hats or the eyes for glasses, proved particularly challenging. The original model's masks were optimized for facial expressions, and modifying them to include non-facial regions introduced significant complexity and often disrupted the model's overall performance. Additionally, the lack of annotated datasets for accessories further complicated the training process, as existing datasets provided limited coverage for these attributes. Despite exploring the use of synthetic data to simulate accessory attributes and introducing additional loss functions to enforce accessory realism, these measures were insufficient to fully address the challenges. Ultimately, the time and effort spent highlighted the fundamental incompatibilities between the AU-driven framework and the requirements of accessory generation, preventing the successful integration of this functionality within the GANimation architecture.

7 Conclusion

In this paper, we present an extension of the GANimation framework to incorporate age progression and regression into anatomically aware facial animation. Our methodology introduces age-related parameters alongside Action Unit (AU) conditioning, attempting to blend aging transformations with dynamic facial expressions. But it was found that while aging and dynamic facial expressions worked in isolation, in conjunction the results were sub-par.

Despite challenges with integrating additional features, such as gender and accessory transformations, our approach underscores the feasibility of joint age-expression modeling in GAN-based architectures. Future research can explore the extension of this framework to address limitations and incorporate a wider set of features, such as changing the background of an image or altering the lighting. Additionally, further work can be done to fine-tune the GANimation model with age to receive better results than presented, and there is also the possibility of applying similar age conditioning to video.

References

Cross-age reference coding for age-invariant face recognition and retrieval. https://bcsiriuschen.github.io/CARC/. [Online; accessed 15-November-2024].

Cpapers with code—ffhq dataset. https://paperswithcode.com/dataset/ffhq. [Online; accessed 15-November-2024].

Yang J. Xue N. Kotsia I. & Zafeiriou S. Deng J., Guo J. Arcface: additive angular margin loss for deep face recognition, 2022.

Sindagi V. A. Patel V. M. Di, X. Gp-gan: gender preserving gan for synthesizing faces from landmarks, 2018.

Roldan F. Tubau M. Escur J. Pascual S. Salvador A. Mohedano E. McGuinness K. Torres J. Giro-i-Nieto X. Duarte, A. Wav2pix: Speech-conditioned face generation using generative adversarial networks, 2019.

Aila T. Laine S. Lehtinen J. Karras, T. Progressive growing of gans for improved quality, stability, and variation, 2018.

Bao J. Yang H. Chen D. Wen F. Li, L. Faceshifter: Towards high fidelity and occlusion aware face swapping, 2020.

Zhu Y. Wang Y. Lin C.-W. Ghanem B. & Shen L. Li, B. Anigan: Style-guided generative adversarial networks for unsupervised anime face generation, 2021.

Agudo A. Martinez A. M. Sanfeliu-A. Moreno-Noguer F. Pumarola, A. Ganimation: Anatomically-aware facial animation from a single image, 2018.

Gu J. Tang X. Zhou-B. Shen, Y. Interpreting the latent space of gans for semantic face editing, 2020.

Newson A. Gousseau Y.-& Hellier P. Yao X., Puy G. High resolution face age editing, 2020.