Facial Expression Synthesis

Abdurrehman Syed
University of Oulu
abdurrehman.syed@student.oulu.fi

Abstract

This report summarizes four papers to highlight current research in facial expression synthesis. Discusses two key aspects of static and dynamic synthesis, their motivation and challenges and explores and evaluates various GAN techniques for image and video generation, evaluating results and mentioning future works.

1 Introduction

Facial expressions serve as a universal means of human communication and perception. In recent years, the idea of computers understanding and synthesizing these expressions has evolved the field of facial expression synthesis. Expanding its use cases to various domains, such as Virtual Reality (VR), graphic designing, animations, entertainment, and many more. This evolution has been largely driven by advancements in the field of artificial intelligence, particularly the emergence of Generative Adversarial Networks (GANs). GANs are a class of deep neural networks that are very powerful in image generation tasks and self-learning capabilities. The adversarial training process enables them to autonomously improve over time, enhancing their ability to create more photo-realistic and convincing images as they continue to refine their internal parameters. From an abstract point of view, GANs consist of two components, a generator, which keeps on generating image data and a discriminator, which continuously categorizes that data to be either real or fake. After each iteration the loss is calculated and accordingly the weights are shifted for both components, proving this two-player minmax architecture to be a work of art. Facial expression synthesis has gained substantial attention after the development of GANs, however, there are still certain challenges and limitations that need to be addressed in-order to generate high quality facial expressions. For instance, the lack of realistic expression, poor retention of facial identity features, low synthesis efficiency is still proving to be a hindrance in static expression synthesis. Similarly, for dynamic expression synthesis where the goal is to create a realistic video from a single image with accurate landmark mapping while preserving facial and temporal data can be an even greater challenge. To address these complexities the researchers have further evolved the concept of GANs to more complex forms including, Conditional GANs, Wasserstein GANs, starGANs, and their different types. Current research is continuously tuning and modifying these architectures to address the challenges mentioned above. In this report, we will cover four papers which solve some of the issues mentioned above.

2 Related Work

Recently, we see increase in research for static and dynamic facial expression synthesis within the broader context of facial expression generation. So, we divided this section into two corresponding parts which cover the motivations, problems, and solutions of two papers each from static and dynamic synthesis literature. In the rest of this report, paper [1] will be referred as cGAN-P1, [4] as LGPGAN-P2, [2] as cWGAN-P3, and [3] as GANimation-P4.

2.1 Static Facial Expression Synthesis

Facial expression transfer methods primarily fall into two categories. The first category relies on traditional 3D modeling to construct facial parameter models for aligning target faces with source facial expressions, often using mesh fitting and 3D morphable models (3DMM). However, this method requires the design of specific facial model parameters for each character, limiting generalization and control over facial details that may result in overlooking the preservation of non-key target facial features. The other category utilizes deep learning, particularly GANs, to directly generate target face images with desired source expressions. GANs have significantly improved image quality and exhibit a high fitting ability in image generation. They have excelled in the field of facial attribute editing, with models like StarGAN and AttGAN setting the benchmark. These methods allow for changing facial attributes effectively, and they serve as a base for the exploration of facial expression synthesis. Similarly in facial expression transfer, researchers have explored various GAN-based approaches like G2-GAN which employs facial geometric information as a conditional vector to guide expression synthesis. However, it has a limitation as it requires a neutral expression image as an intermediary for expression transfer. Another such approach is ExprGAN which focuses on controlling expression intensity through one-hot vectors and expression codes. Nevertheless, these methods mostly focus on transferring expressions on the face, while often ignoring identity features of the target face. To address these limitations, [1] proposes a novel facial expression transfer model based on cGAN, comprising two components: Facial feature point fusion and expression transfer modules, their combination resulting in a highly realistic generation of the target image with expressions of the base image. In parallel to [1] another model has been proposed known as Local Global Perception GAN (LGP-GAN) in [4], that highlights and solves a key problem in GAN-based traditional approaches that they often treat the whole face as a uniform entity. This approach leads to issues like overlapping and blurring in local facial regions, overlooking the fact that differences between facial expressions are often concentrated in crucial regions, such as the eyes and mouth. Furthermore, research in physiology and psychology indicates the significance of local facial regions in capturing facial expressions accurately. Existing methods, despite their success in facial attribute editing, are not suitable for generalizing the complex task of facial expression synthesis due to the variability in facial deformations. In response to these challenges, the LGP-GAN model works optimally, characterized by a two-stage cascaded architecture. This architecture differs from previous works in that it divides the facial expression synthesis process into two parts: local facial region generation and global facial image generation. By focusing on both local and global facial information, the LGP-GAN can synthesize facial expressions step by step, thus addressing the limitations of prior methods.

2.2 Dynamic Facial Expression Synthesis

While static facial expression synthesis has played a crucial role in modifying facial attributes within images, dynamic facial expression synthesis introduces a higher level of complexity. It involves the task of animating facial expressions across multiple frames while maintaining the integrity of facial features and ensuring smooth temporal and spatial transitions. However, traditional GAN approaches face several challenges when applied to this dynamic context. These challenges include the use of discrete and limited expressions which do not capture the full range of facial movements, smooth flow, adapting GANs and cost functions to work with manifold-valued data, particularly for facial expression generation. To address these limitations, our research explores two innovative solutions, one based on Conditional Wasserstein GAN [2] and second being GANimation [3]. The first mentioned method introduces a CWGAN approach that operates on a Hilbert hypersphere. This model generates compact motion data on the hypersphere, addressing the issue of motion artifacts and enabling smoother motion transitions. It leverages geometric representations of facial landmarks for dynamic expression generation, allowing the generation of various expressions and video sequences with improved image quality while the second idea proposes an anatomically-aware GAN model and is influenced by continuous movement of facial muscles. In contrast to traditional discrete models, this allows generation of realistic facial expression ensuring smooth frame transitions. Since training is unsupervised; it does not need image pairs containing varying expressions.

3 Methodologies

This section discusses the methodologies used in the four static and dynamic facial synthesis papers.

3.1 Static Models

3.1.1 cGAN-P1 [1]

The architecture discussed will be of two-staged based cGAN model, this model consists of two modules: Facial Feature Point Fusion Module and Expression Transfer Module, as illustrated in 1. These modules work in sync to transfer facial expressions from a source image to a target image.

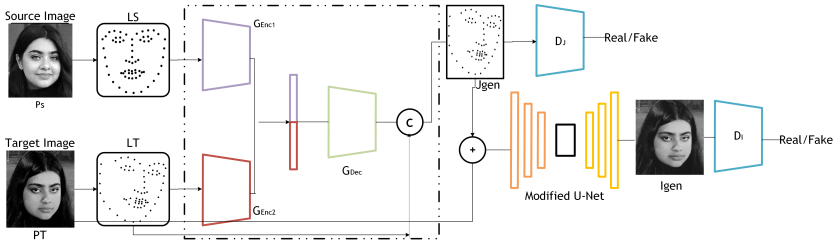

Figure 1: Autoencoder - Conditional GAN

Facial Feature Transfer Module: This module functions as an autoencoder network, initially it extracts spatial information of key facial landmark features using Dlib module (Ls and Lt). Afterwards, this higher dimensional feature data is passed to the encoders (Genc1 and Genc2), these encoders convert the complex features into simpler form, cascades them and then passes this output to the second component of autoencoder, decoder. Along with the output of encoder, the target landmarks are also passed to the decoder. The decoder after receiving the input vectors generates a deviation vector which instructs on how to adjust the landmark features over the target image, the target landmarks are used as a canvas where the decoder applies the adjusted key feature vectors After which the result is passed forward to the decoder and the next module of the model.

Expression Transfer Module: This module is a modified U-net network, specifically designed for facial expression transfer. It undergoes specific modifications based on DCGAN principles to improve spatial up and down-sampling, address pixel value ranges, and prevent checkerboard artifacts. This modified U-net network consists of 3 primary structures: up-sampling layers, down-sampling layers, and bottleneck layers, each with their specific activation functions and convolutional operations. It receives the generated face feature key point image (Jgen) and the target image (Pt). The U-net structure facilitates the transfer of facial expressions while retaining most face features and background, this output is then passed on to the markov discriminator for assessment. This discriminator employs convolutional layers and spectral normalization to ensure training stability and categorizes output image as real or fake. Upon completion of these processes, the model successfully transfers facial key features from source to the target while preserving essential characteristics.

3.1.2 LGPGAN-P2 [4]

The proposed LGP-GAN is a two-stage cascaded architecture designed for facial expression synthesis. It aims to generate a target facial expression while retaining identity properties.
This model consists of two local networks which capture the local texture details of the eye and mouth region separately while one global network that learns the whole facial information. This design is motivated by the fact that different facial expressions often occur in crucial regions (eyes and mouth). Existing methods sometimes overlook local regions, leading to overlapping and blur. The cascaded structure helps the network better learn features in different local regions of the face.

The Local regions are identified based on the Facial Action Coding System (FACS), which defines 44 Action Units (AUs) representing basic muscle motions. Most AUs relevant to facial expressions are found in the eye and mouth regions due to which facial landmarks are used to track these key regions.

The GAN in this model consists of 3 generators two locals $(D_{eye} and D_{mouth})$ and one global (D_{global}). Although each of these generators have similar architectures, they are assigned different tasks. Attention mechanisms are employed to focus on areas that need to accommodate changes during facial expression synthesis and the final output of the generator incorporates both color and attention masks to control the synthesis process. It's calculated adaptively during generation.

Similarly, the discriminator is designed for each stage and for global evaluation. It contains two local discriminators $(D_{eye} and D_{mouth})$ and one global discriminator (D_{global}). Local discriminators eval-

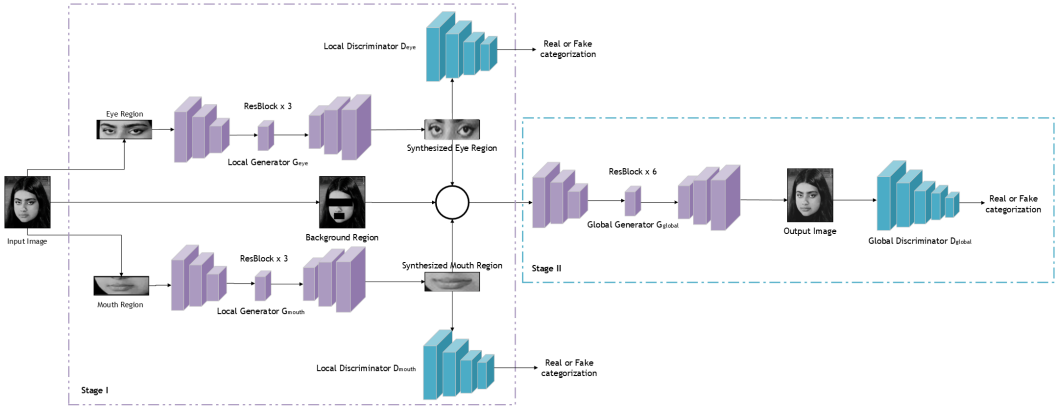

Figure 2: Local Global Perception GAN

uate the quality of local details in different regions, while the global discriminator assesses combined facial features. They employ a PatchGAN model to distinguish between real and synthesized facial expressions and predict the category of the facial expression either to be real or fake.

This approach enhances the synthesis quality, focusing on the most informative regions while avoiding the generation of blurred or unrealistic local details making it valuable for applications in various fields, including healthcare and entertainment.

3.2 Dynamic Models

3.2.1 cWGAN-P3 [2]

DC-WGAN model is designed to generate facial expression sequence given a neutral face image, its landmarks, and desired expression. It decomposes the problem into two functions, one that learns the distribution of facial expression dynamics (motion) and another that focuses on synthesizing texture information. This decomposition allows for efficient video generation and the ability to apply generated dynamics to different identities or generate videos of the same person with different facial expressions. The model comprises two main components: MotionGAN and TextureGAN.

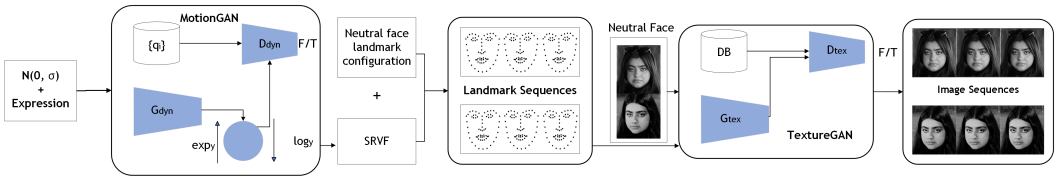

Figure 3: Wasserstein cGAN

MotionGAN solely focuses on facial landmarks and expression dynamics from the given input images. It treats those landmark sequences as points on a hypersphere manifold and uses the CW-GAN model on the hypersphere to learn the distribution of expression dynamics. The generator maps random noise to an SRVF point on the hypersphere, while the discriminator distinguishes between real and generated SRVF points. After training, MotionGAN can generate the dynamics of new facial expressions in the form of SRVFs.

Texture GAN, given the generated expression dynamics, generates the static image frames for each expression phase, focusing on adding texture information to the generated expression dynamics. It uses an adversarial model to generate images conditioned on landmarks. The adversarial loss encourages generated images to be indistinguishable from real images and an identity loss ensures that the identity of the face remains consistent between input and output while the reconstruction loss ensures that the generated frames are close to the ground truth. Therefore, learning to generate frames that preserve identity and texture while following the previously learned expression dynamics.

This Motion and TextureGAN combination provides a model for generating expressive and identity-preserving videos of faces performing various facial expressions from a single neutral expression.

3.2.2 GANimation-P4 [3]

This model shares a lot of similarities with CW-GAN model but the goals are quite different. It uses AU annotations and focuses on controlling the magnitude of activation of each AU to generate various expressions, going beyond a fixed set of expressions and allowing more fine-grained control over the generated expressions. Moreover, GANimation introduces a fully unsupervised training approach.

Consisting of two main modules: a generator (G) and a conditional critic (D) based on the WGAN with Gradient Penalty (WGAN-GP). Initially, the model starts with a dataset of facial images where each image is annotated with the activated AUs, describing the specific muscle movements responsible for facial expressions. Afterwards, the annotated image is passed into the generator which includes an attention mechanism which focuses only on regions responsible for synthesizing facial expressions while leaving static elements like hair. Outputting two masks (color mask and an attention mask), the final image is computed as a combination of these masks and the original image. Condition critic (discriminator) based on Patch-GAN evaluates generated images in terms of photo-realism and expression.

The loss functions used in this model include adversarial loss which is based on WGAN-GP for measuring realism of generated images and works better than traditional adversarial loss in GANs, attention loss for smooth attention masks which also prevents it from saturating to extreme values, conditional expression loss, and identity loss for exhibiting desired expressions and preserving identity of the face. These loss functions collectively work to ensure that the generated images are both realistic and expressive, fulfilling the goals of GANimation.

4 Datasets

RaFD (The Radboud Faces Database) comprises 67 participants and a total of 8040 images. It includes 8 emotions: happiness, sadness, surprise, anger, fear, disgust, contempt, and neutrality. Each emotion variant is accompanied by 3 distinct gaze directions, and multiple angles are captured simultaneously using 5 cameras. The papers cGAN-P1 [1], LGP-GAN-P2 [4], and GANimation [3] have used this database. **Extended Cohn-Kanade Dataset (CK+)** dataset comprises 123 participants with a total of 593 facial expression sequences. Each sequence starts with a plain neutral expression and ends with a peak expression, and all of them are annotated using FACS. Most of the images are grayscale, with few color expression sequences, that need to be converted to grayscale before the experiment. The papers cGAN-P1 [1] and cWGAN-P3 [2] use this dataset for training. **Oulu-CASIA** which has over 480 videos, each with 80 different people. For every person, there are six videos showing different emotions. All the videos start with a neutral face and slowly grow to the strongest form of that emotion. cWGAN-P3 uses this dataset [2]. **MUG Facial Expression**, a database that includes videos of 86 subjects. Each video consists of 50-160 frames and fundamental expressions. The start and finish of each video show neutral expressions, therefore, for the models only the first half part from neutral to pinnacle is relevant. cWGAN-P3 uses this dataset [2]. **EmotioNet** is a comprehensive collection of 1 million facial expression images of various emotions. These images are annotated with discrete AUs activations. Allowing researchers to explore facial muscle motion and expression combinations. GANimation [3] uses this image dataset.

5 Experiments and Evaluation

This section highlights the metrics and findings of the models discussed in this report.

The images generated by cGAN-P1 [1] are compared with CGAN, C2GAN and Pix2Pix models and evaluated based on the following two metrics: structure similarity (SSIM) and peak signal to noise ratio (PSNR). Results show that on RaFD dataset the SSIM and PSNR values of the model are better than CGAN and Pix2Pix models. However, on CK+ dataset it exceeds all other models in accuracy.

In LGP-GAN-P2 [4] two metrics are used for quantitative evaluation of the images, Inception Score (IS) and Fréchet Inception Distance (FID). The model is assessed against the state-of-the-art models including StarGAN, AttGAN, and GANimation, the findings revealed that GANimation outperformed AttGAN and StarGAN in facial expression synthesis, highlighting the complexity

of the task due to significant transformations and deformations in facial regions. The study also emphasized that existing methods tend to generate blur and overlap around local regions during large facial deformations, as they focus on general facial information rather than local facial regions. The proposed LGP-GAN, with its two-stage cascaded architecture, achieved superior overall quality.

For cWGAN-P3 [2] the proposed approach is assessed qualitatively and quantitatively by evaluating the facial expression generated by MotionGAN and the quality of videos generated by TextureGAN. For landmark generation its qualitative results are compared with the actual output landmarks whereas the quantitative results are assessed using Geodesic distance to measure the similarities. Both of them show promising results. In video generation, the quality of videos is compared with standard methods, such as MoCoGAN, VGAN, and TGAN, and for quantitative evaluation several metrics are used, including IS, ACD, and ACD-I. For the next task of identity features visualization where generated videos are visualized in a 2D space using Multidimensional Scaling (MDS) providing insight of how well the model preserves identity.

The paper GANimation-P4 [3] initially tests the ability of its model to edit single AUs with varying intensities while preserving the identity of the person. The evaluation involves individual transformations of nine AUs at four intensity levels. At the 0-intensity level, the model successfully maintains the original AU, ensuring no unwanted changes. For non-zero intensities, the model progressively enhances each AU, showcasing its capability to convincingly render complex facial movements. Further, the model is compared to various baselines, including CycleGAN and StarGAN, in the task of discrete emotions editing. The model's distinctive approach, which focuses on a continuum of expressions and uses an attention mask leveraging it to produce a great combination of detailed expressions with only few AUs, resulting in more expression variability and higher-resolution images.

6 Conclusion, Challenges and Future Works

This report highlights the advancements in static and dynamic field of facial expression synthesis, which have yielded novel frameworks for synthesizing images and videos with remarkable precision. These developments show potential for a wide range of applications in various domains. The models discussed give a glimpse of what GANs are capable of, their evolution and different approaches researchers have used to improve their output. Each model presented has its own challenges, limitations that they address and, architectural solution. We explored the cGAN-P1 [1] using landmark detection, giving realistic outputs, the approach of separating local and global features [4] giving more robust detail and texture to the expressions on the target image. Examined dynamic facial synthesis using cW-GAN-P3 [2] and GANimation-P4 [3] models innovatively using AUs for their unsupervised training in a continuous space. However, these developing models have various limitations, such as non-human input images, the challenge of outputting more than basic expressions that we currently are getting, smoother spatial and temporal transitions, overfitting and cross validation for generators, calculating proper loss, model's limited performance in uncontrolled conditions, and discrete expression generation are some of the challenges that are being faced on making the model more general. In future we can extend GANimation to video sequences, advancing from expression analysis in 2D images to 3D frames, exploring 3D facial expression and action generation, and expanding expression synthesis to unconditional and real-world scenarios.

References

[1] Yang Fan et al. "Facial Expression Transfer Based on Conditional Generative Adversarial Networks". In: *IEEE Access* 11 (2023), pp. 82276–82283.

[2] Naima Otberdout et al. "Dynamic Facial Expression Generation on Hilbert Hypersphere With Conditional Wasserstein Generative Adversarial Nets". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.2 (2022). DOI: 10.1109/TPAMI.2020.3002500.

[3] Albert Pumarola et al. "Ganimation: Anatomically-aware facial animation from a single image". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 818–833.

[4] Yifan Xia et al. "Local and Global Perception Generative Adversarial Network for Facial Expression Synthesis". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.3 (2022), pp. 1443–1452. DOI: 10.1109/TCSVT.2021.3074032.