SuperGAN: Face Swapping and Reenactment with Improved Face Restoration

Anant Singh
New York University
anant.singh@nyu.edu

Shantanu Kumar
New York University
sk9698@nyu.edu

Stuti Biyani
New York University
sb7580@nyu.edu

Abstract

In this work, we propose a General Adversarial Network that improves the results of Face Swapping and Reenactment by improving the resolution and quality using face restoration techniques. We first started by designing a stacked model which uses the results of FSGAN [2] as input to GFPGAN [3] which then using face restoration technique improves resolution and quality. Then based on our understanding of both the models, we designed an Integrated GAN architecture: **SuperGAN**. SuperGAN crops the face from the source and target, blend the target with source and finally upscale the image and improve pixel level quality of the image. Finally the swapped face is merged with the original background. This process is repeated for all the frames in the video and we were able to achieve face swapping and reenactment with improved facial feature resolution.

1 Introduction

Face swapping is the practice of transferring a face from one person to another, source image to target image, such that a face is smoothly replaced appearing in the goal and yielding a plausible outcome. It also involves *Face reenactment* which is a technique that relies on a control face in one video to direct the motions, expressions and deformations of a face appearing in another video or picture. Due to their applications in entertainment [8] privacy [9] and training data creation, these tasks are gaining a lot of study attention.

Using underlying 3D face representations [10] to transfer the face look, most recent publications [6], [7] suggested methods for either swapping or reenactment, but seldom both. Face forms were either approximated or fixed based on the supplied image. The 3D form was then aligned with the input photos and utilized as a proxy to swap or modify the face expression and viewpoint.

Face swapping and reenactment were also done using deep networks. Generative Adversarial Networks (GANs) [11] were found to be capable of producing realistic fake face pictures. Multiple face reenactment approaches were motivated using conditional GANs [12] to shift a picture showing genuine data from one domain to the other. The Deepfakes [7] project used convolutional neural networks to do face swapping in films, making switching broadly accessible to non-experts and garnering widespread media attention. By replacing the traditional graphics pipeline with implicit 3D face representations, such approaches can create more realistic face pictures.

To dissociate the identity component of a face from other qualities such as position and expression, some approaches used domain separation [13] in latent feature spaces. The identification is represented as the manifestation of latent feature vectors, which results in severe information loss and lowers the image quality. Topic-specific techniques are specifically trained for each swapped or reenacted subject or set of subjects. As a result, considerable training sets per subject are required to attain respectable results, which limits their potential use. A key issue with earlier face generation approaches, particularly 3D-based methods, is that they all demand extra caution when dealing with partly occluded faces.

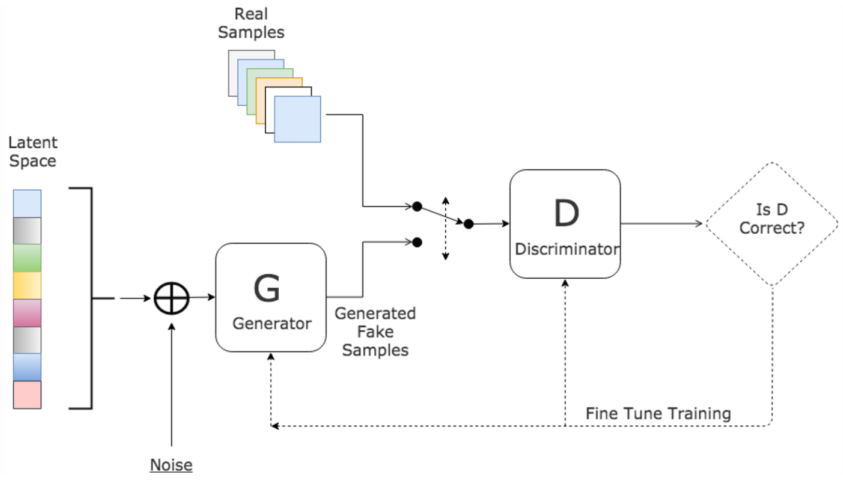

Figure 1
Generative Adversarial Network

A new model Face Swapping GAN (FSGAN) [1], [2] presents a method for interpolating between facial landmarks such as pose, expression, and identity without relying on 3D data, inpainting generator is improved by adding symmetric and face landmark cues, a post-process step is added to reduce the jittering and artifacts and added a new metric for comparing expressions. Also, two additional losses are proposed: a step-wise consistency loss for gradually teaching face reenactment in tiny increments, and a Poisson blending loss for training the face blending network to smoothly merge the source face into its new environment.

Although the output generated by FSGAN avoids time-consuming, subject-specific data collecting and model training, allowing non-experts to do face swapping and reenactment, it still has its limitation. Analyzing reenactment outcomes for various facial yaw angles the higher the angular discrepancies, the worse the quality of the identity and texture. Furthermore, using the face reenactment generator too many times blurs the texture. Because these low-quality inputs are unable to provide precise geometric priors and high-quality references are unavailable, the application in real-world settings is limited. There is no penalty for identity loss and facial features don't have much importance in terms of loss. These drawbacks of FSGAN makes output blurry and with missing facial details.

To address these issues, we propose SuperGAN. SuperGAN improves the precision and resolution of outputs using a Restoration Generator based on GFPGAN. For face restoration, we use a wide set of generative facial priors which provide enough facial texture and color information for us to execute face restoration and color enhancement together. To include generative facial prior, we used the GFP-GAN framework with careful designs of structures and losses. In a single forward pass, SuperGAN with Channel-Split Spatial Feature Transform (CS-FST) layers provides a decent mix of fidelity and texture faithfulness.

2 Methodology

2.1 Background

In this section we describe in detail the FSGAN and GFPGAN models. Both these models are based on GAN. GAN or Generative Adversarial Networks is an approach to generative modeling using deep learning methods. It is an unsupervised learning task in machine learning that involves automatically discovering and learning the regularities or patterns in input data in such a way that the model can be used to generate new examples that could have been drawn from the original dataset. GANs can be broken down into three parts [4]:

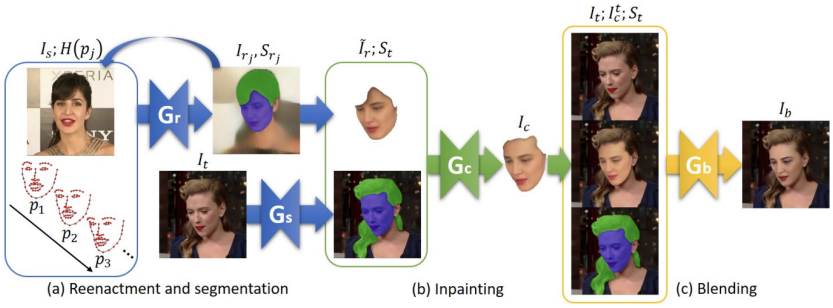

Figure 2
FSGAN
Source: FSGANv2: Improved Subject Agnostic Face Swapping and Reenactment [2]

- **Generative:** To learn a generative model which describes how data is generated in terms of a probabilistic model.
- **Adversarial:** The training of a model is done in an adversarial setting.
- **Networks:** Use deep neural networks as the artificial intelligence (AI) algorithms for training purpose.

GANs consist of a **Generator** and **Discriminator**. The generator basically generates fake data samples and tries to fool the discriminator. While the discriminator tries to distinguish between the real and fake samples as depicted in Fig 1.

2.1.1 FSGAN

Face Swapping is the transformation of a face from a source image to a target image. The resulting image replaces the face in the target image. *Face Reenactment* is a conditional face synthesis task that aims at fulfilling two goals simultaneously: 1) transfer a source face shape to a target face; while 2) preserving the appearance and the identity of the target face [5].

FSGAN is basically Face Swapping and Reenactment GAN which was proposed by Nirkin et al., first in 2019 and then later improved by them in 2022. Their latest work on FSGAN offers a subject agnostic swapping scheme that is applied to pairs of faces without training on those faces. This is depicted in Fig 2.

FSGAN consists of three major components:

- *Reenactment Generator* which estimates the reenacted face, and the *segmentation generator* that estimates the face and hair segmentation mask of the target image.
- *Inpainting Generator* inpaints the missing parts of the reenacted face based on the target to estimate the complete reenacted face.
- *Blending generator* blends the complete reenacted face and the target face using the segmentation mask.

2.1.2 GFPGAN

Generative Facial Prior GAN is a GAN architecture is designed to upscale the quality of human faces in damaged and low resolution photos. A GFPGAN works in the following way:

- First, a *degradation removal module* takes the photo and removes degradation. It extracts 2 types of features: the latent features to map the input image to the latest StyleGAN2 code and multi-resolution spatial features for modulating the StyleGAN2 features [3].
- Then a pretrained StyleGAN2 model acts as a generative facial prior. In this stage, intermediate convolutional features are produces with an intent of using spatial features to modulate the final output.

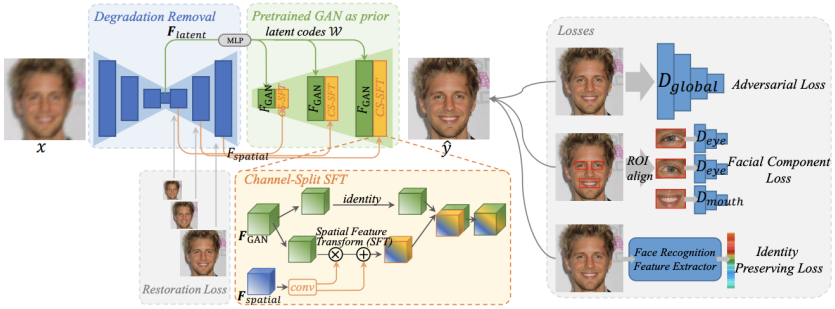

Figure 3
GFPGAN
Source: Towards Real-World Blind Face Restoration with Generative Facial Prior [3]

- The *Channel-Split Feature Transform* allows the spatial features to be used to predict the transform parameters. These parameters can be used to scale and displace the features in the feature maps in the generator.
- And finally, GFPGAN radically restores and upscales the quality of the faces of the input images.

2.2 Architecture

We have integrated FSGAN and GFPGAN to improve the results of Face Swapping and Reenactment in terms of image or video quality. We started by implementing a simple stacked model that takes the output of FSGAN, splits it into frames, and then passes it to GFPGAN which then improves its quality. After understanding how both (GFPGAN and FSGAN) function, we designed a state-of-art architecture, we call the *SuperGAN*. SuperGAN is a single GAN architecture that can be used for Face Swapping and Reenactment with a significant improvement in the quality and resolution over the previous works. Now we will discuss both the architectures in detail.

2.2.1 A Stacked Model

Our first intuition was to explore the models individually and understand the inner working of the two models - FSGAN and GFPGAN. So we run the models and generated sample outputs by stacking the models in a very basic sense. We started with configuring both the models separately and make the necessary changes to run a stacked model on NYU Greene. The workflow of this architecture is as follows 4:

1. Two videos - source and target, are provided to the FSGAN module as inputs. The output is a video with the target face swapped with the face in the source video.
2. Because GFPGAN only allows images as inputs, we split this output video into frames. All these frames are then provided as input to the GFPGAN module. This module helps restore the face and improve the quality. The output is all the frames with enhanced quality.
3. Finally, we combine all the frames and merge it with the original audio to get the swapped video with better resolution and quality.

Drawbacks:
This architecture gives very good results in terms of quality and resolution. But in this case, the video has to go through three different modules - FSGAN, Splitting the frames, and GFPGAN. Hence, it increases the time of processing, and introduces unnecessary repetition of functions. This stacked version can't fully utilise the concepts introduced in GFPGAN.

We overcome these drawbacks by implementing our second architecture which is developed by integrating both the models into a single GAN - called SuperGAN. This is discussed in detail in the next section.

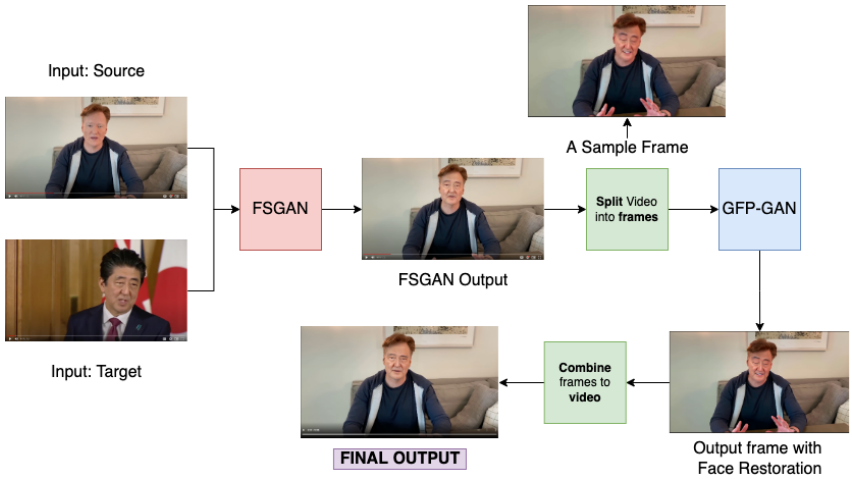

Figure 4
Architecture 1: Stacked GAN Model

2.2.2 SuperGAN

In this architecture, we introduce a new generator after the blending and completion generator in FSGAN. So once the blending generator from FSGAN blends the swapped image with the segmentation mask, the result tensor is passed to our *Restoration Generator* G_{rs}. This restoration generator uses concepts from GFPGAN to upscale and enhance facial features of the result tensor from FSGAN. G_{rs} acts as an enhancer module, that helps gain more details in facial features.

The workflow of SuperGAN is as follows (Figure 5):

1. The source and target videos are provided as input to the Reenactment and segmentation generators. The reenactment generator outputs the reenacted face, and the segmentation generator outputs the segmentation mask of the target face.
2. Then the inpainting generator inpaints the missing parts of the reenacted face based on the segmentation mask, and outputs a completed reenacted face.
3. Then this reenacted face and the target face are blended by the blending generator, and completed with the completion generator.
4. Finally, this result tensor is used as input for our restoration generator. This generator first uses a degradation removal module, and the result is passed to a pre-trained face GAN as facial prior.

2.3 Training Losses

2.3.1 Domain Specific Perceptual Loss

We employ perceptual loss to capture tiny facial characteristics, which has been widely used in recent work for face synthesis, outdoor sceneries, and super resolution. Perceptual loss compares high frequency information using a Euclidean distance utilizing the feature maps of a pretrained VGG network. The Perceptual Loss is give by:

$$\mathcal{L}_{perc}(x, y) = \sum_{i=1}^{n} \frac{1}{C_i H_i W_i} ||F_i(x) - F_i(y)||_1 \tag{1}$$

where C_i is the number of channels, and H_i, W_i are the height and width dimensions.

Using a network pretrained on a generic dataset like ImageNet, it is difficult to properly capture the subtleties inherent in face photos. Instead, the network is pre-trained on VGG-19 networks for face recognition and attribute categorization[1].

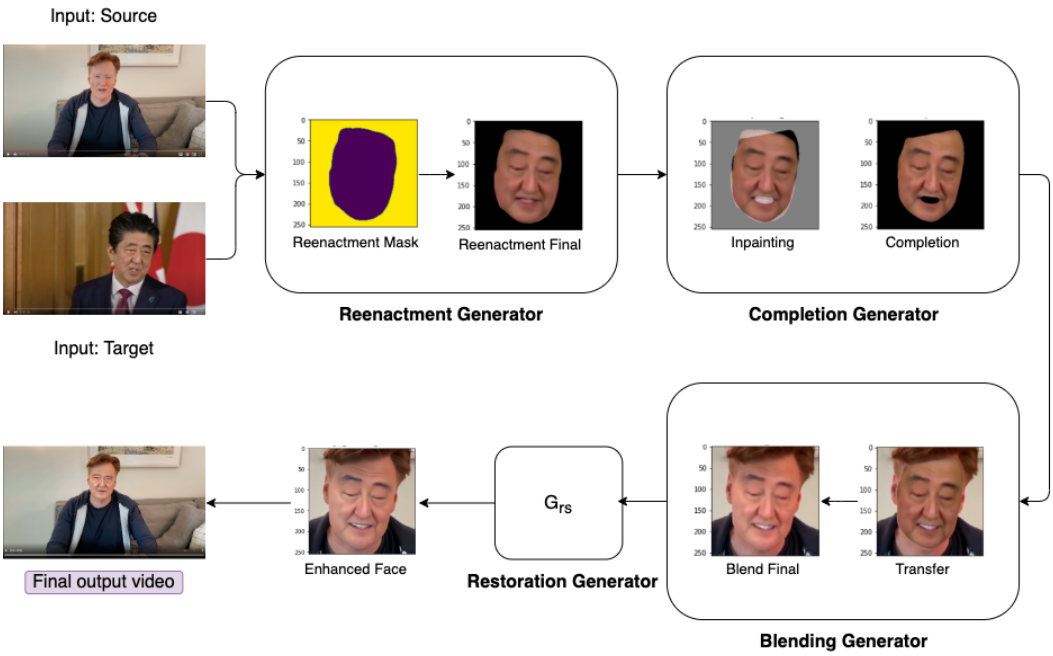

Figure 5
Architecture 2: SuperGAN

2.3.2 Reconstruction Loss

While Equation 1's perceptual loss retains small details effectively, generators trained solely on that loss frequently create pictures with incorrect colors, owing to reconstruction of low frequency visual material[1].

$$\mathcal{L}_{pixel}(x, y) = ||x - y||_1 \tag{2}$$

The overall loss is given by:

$$\mathcal{L}_{rec}(x, y) = \lambda_{perc}\mathcal{L}_{perc}(x, y) + \lambda_{pixel}\mathcal{L}_{pixel}(x, y) \tag{3}$$

2.3.3 Adversarial Loss

An adversarial objective is used to increase the realism of our produced pictures. We used a multi-scale discriminator made up of many discriminators, $D_1, D_2, ..., D_n$, each of which operated at a distinct picture resolution[1].

$$\mathcal{L}_{adv}(G, D) = \min_{G} \max_{D_1, D_2, ..., D_n} \sum_{i=1}^{n} \mathcal{L}_{GAN}(G, D_i) \tag{4}$$

where $\mathcal{L}_{GAN}(G, D)$ is defined as:

$$\mathcal{L}_{GAN}(G, D) = E_{(x,y)}[logD(x, y)] + E_x[log(1 - D(x, G(x)))]. \tag{5}$$

2.3.4 Facial Component Loss

Facial component loss is also included with local discriminators for the left eye, right eye, and mouth to improve the perceptually meaningful face components even more. As illustrated in Figure 3, first ROI is used to align and crop interesting regions. Unlike earlier feature matching losses with spatial limitations, Gram matrix statistics of real and restored patches are sought to use.

Gram matrix computes feature correlations and, in most cases, successfully captures texture information. Features are extracted from the taught local discriminators' many layers and learn to match these Gram statistics of intermediate representations from the real and restored patches.

In terms of providing realistic face details and eliminating unpleasant artifacts, it was discovered that the feature style loss outperforms the prior feature matching loss empirically[3].

$$\mathcal{L}_{comp} = \sum_{ROI} \lambda_{local} E_{\hat{y}_{ROI}}[log(1 - D_{ROI}(\hat{y}_{ROI})] + \lambda_{fs} ||Gram(\psi(\hat{y}_{ROI})) - Gram(\psi(y_{ROI}))||_1$$

where ROI is region of interest, i.e., left eye, right eye, mouth. D_{ROI} is the local discriminator for each region. ψ denotes the multi-resolution features, and λ_{local} and λ_{fs} represent the loss weights of local discriminative loss and feature style loss respectively.

2.3.5 Identity Preservation Loss

In the model, we use identity preserving loss. We define the loss based on the feature embedding of an input face, similarly to perceptual loss. The pretrained facial recognition ArcFace model is used, which captures the most important aspects for identification discrimination[3].

$$\mathcal{L}_{id} = \lambda_{id} ||\eta(\hat{y} - \eta(y))||_1 \tag{6}$$

where η represents the face feature extractor.

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \mathcal{L}_{adv} + \mathcal{L}_{comp} + \mathcal{L}_{id} \tag{7}$$

2.4 Challenges:

Due to the large size of the FSGAN model it had become difficult to understand the working of the model. The code repository is very extensive and complex to understand. So our main challenge was to figure out how to include the face restoration process with the blending and completion generator of FSGAN. Another challenge was restructuring the model. This process was challenging because of the complicated FSGAN architecture.

3 Experimental Results

We performed experiments to verify our proposed framework. We compare our method with the results of FSGAN. Figure 6 shows frames from one of our experiments. It represents the source and target videos, and comparison of results with FSGAN and SuperGAN. As can be seen from the figure, SuperGAN significantly improves the resolution and quality of the results as compared to the previous work in this field. Figure 7 shows the effect of using Restoration generator on FSGAN.

Figure 6
Comparison of FSGAN with SuperGAN

Figure 7
Before and After using Restoration Generator

4 Conclusion

We have proposed the SuperGAN framework that leverages rich and diverse facial prior with Face Swapping and Reenactment. We add a new restoration generator based on GFPGAN in FSGAN. Our method helps gain detailed facial features which improves the resolution of FSGAN output, thus making it more realistic. Our comparison demonstrates the superior capability of SuperGAN in Face Swapping and Reenactment over previous work. SuperGAN maintains the approach of subject agnostic face swapping, which means that it can be applied to faces of different subjects without requiring subject specific training.

Our Contribution: FSGAN is used to swap faces given source and target images or videos. But as per the latest work of Nirkin et al. [2], there was still room for improvement with respect to the quality and resolution of the output. We improved the quality by employing concepts from a Face restoration framework GFPGAN proposed by Wang et al. [3]. We proposed SuperGAN, a framework that is used for Face Swapping and Reenactment with an additional Restoration Generator. SuperGAN improves the quality and resolution of FSGAN which produces realistic results. Our code can be found at https://github.com/95anantsingh/NYU-SuperGAN.

References

[1] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject Agnostic Face Swapping and Reenactment, 2019

[2] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGANv2: Improved Subject Agnostic Face Swapping and Reenactment, 2022

[3] Xintao Wang, Yu Li, Honglun Zhang, Ying Shan. Towards Real-World Blind Face Restoration with Generative Facial Prior, 2021

[4] https://zhongpeixiang.github.io/generative-adversarial-network-overview/

[5] Yunxuan Zhang, Siwei Zhang, Yue He, Cheng Li, Chen Change Loy, Ziwei Liu. One-shot Face Reenactment, 2019

[6] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), pages 67–74. IEEE, 2018.

[7] DeepFakes. FaceSwap. https://github.com/deepfakes/faceswap. Accessed: 2019-02-06.

[8] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. Creating a photoreal digital actor: The digital emily project. In Conf. Visual Media Production, pages 176–187. IEEE, 2009.

[9] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging faces in images. Comput. Graphics Forum, 23(3):669–676, 2004.

[10] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gerard Medioni. Extreme 3D face reconstruc- ´ tion: Looking past occlusions. In Proc. Conf. Comput. Vision Pattern Recognition, 2018.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[12] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[13] Yu Tian, Xi Peng, Long Zhao, Shaoting Zhang, and Dimitris N Metaxas. Cr-gan: learning complete representations for multi-view generation. arXiv preprint arXiv:1806.11191, 2018.