Controlling Face's Frame generation in StyleGAN's latent space operations

Agustín Roca[a], Nicolás Ignacio Britos[a]

[a]*Instituto Tecnológico de Buenos Aires, Ciudad Autónoma de Buenos Aires, Argentina*

Abstract

Previous research has suggested that human recognition of faces is heavily influenced by contextual features, such as the outer contour of the face, ears, chin, and hairline, collectively known as the face frame. However, the relationship between the latent space of StyleGAN and the outer contour of the generated faces has received little attention so far. In this paper, we explore the latent space of StyleGAN to preserve the face frame when projecting an image to the latent space or moving through latent directions. Our method involves a post-processing step that modifies the generated images to enhance the preservation of the face frame. Our results show that our method is effective in preserving the face frame with the tested images.

Keywords: StyleGAN, Generative adversarial network, Latent space, Face frame, Image editing, Human perception, Latent directions, Projection

1. Introduction

Recognizing faces is a fundamental ability for humans, but it's a complex task that depends on various factors. According to several studies, features such as the outer contour of the face, ears, chin, and hairline can significantly affect how we perceive and remember a person's face [1] . These set of features, is what we refer to as the face frame.

Email addresses: aroca@itba.edu.ar (Agustín Roca), nbritos@itba.edu.ar (Nicolás Ignacio Britos)

Preserving the face frame has important implications for various face editing applications, such as face swapping, face aging, and virtual makeup. In these applications, the goal is to modify the appearance of a face while keeping its identity intact, or to generate a new face that looks similar to a given face. The face frame plays a crucial role in these tasks, as it provides a stable reference for the overall shape and proportions of the face. If the face frame is lost or distorted during the editing process, the resulting face may look unrealistic or unrecognizable, which can limit the usability and appeal of the application. By preserving the face frame when generating or manipulating faces, we can ensure that the edited faces retain their identity and resemblance to the original faces, which can enhance the user experience and the realism of the results.

There are many image editors applications that are investing in artificial intelligence to achieve better results in their editing tools. However, despite their importance, the relationship between the face frame and the latent space of StyleGAN [2] [3], a state-of-the-art generative model for creating realistic images, has received little attention so far. In this paper, we aim to address this by exploring the latent space of StyleGAN to preserve the face frame when manipulating the generated images. Specifically, we propose a post-processing step that modifies the generated images to enhance the preservation of the face frame.

To do this, we will take advantage of existing neural networks that segments images in three classes: face, hair and background, and use its output to measure the difference of face frame between two images. Using this measurement as a loss function, an algorithm can move through the latent space minimizing the face frame difference.

We analyze how effective is the proposed correction of the face frame for the projection operation and when moving through latent directions. The results show that the method can reduce the face frame difference by a considerable amount.

The paper is organised as follows. Section 2.1 describes how we classify the pixels of the image into hair, face and background. Section 2.2 defines the

function that the algorithm use to measure the face frame difference between two images. Section 2.3 describes the post-processing step that preserves the face frame of the original image. The results and their discussion are presented in Section 3. Section 4 concludes the paper.

2. Materials and Methods

2.1. Face frame difference measurement

There is no standardized way of defining a face frame. We define it as the set of characteristics of a face that gives context for the internal characteristics of such face. A face frame includes the contour of the face, hair, ears. A face frame does not include eyes, mouth, nose or eyebrows. For example, Figure 1 shows two faces with the same frame.

Figure 1: Two different faces with the same face frame.

Having this in mind, we measure the variation of the face frame between two images based on the position of the face and hair in the image.

2.1.1. Image segmentation

The first thing we need to do is to identify the face and hair inside an arbitrary image. This consists of a simple segmentation problem, separate the pixels of an image in three groups: face, hair and background. For this, an open-source pretrained neural network is used [4]. This network proved to give satisfying results for different faces and images with different conditions. This is

essential for the measurement because there is no need to modify any parameters for different images. In Figure 2, we can see some examples of segmentation made by this network. In yellow the pixels classified as hair; in blue the pixels classified as face; and in purple the pixels classified as background.

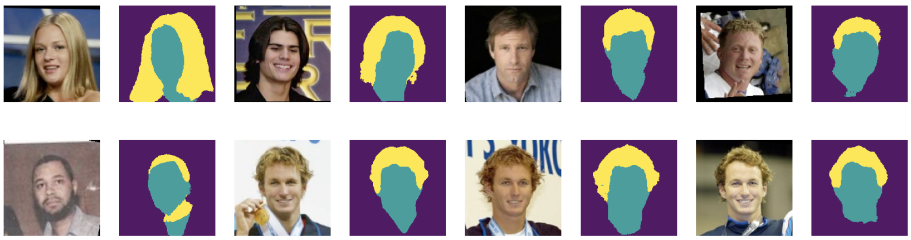


Figure 2: Segmentation made by the neural network. [4]

2.2. Face frame difference formula

Once each pixel of each of the two images is classified, we compare the classifications between the pixels in the same position of the two images. Let f be a function that compares the classifications of two pixels,

$$f(c_1, c_2) = \begin{cases} 0 & c_1 = c_2 \\ 0.2 & c_1 = "Face" \wedge c_2 = "Hair" \\ 0.2 & c_1 = "Hair" \wedge c_2 = "Face" \\ 1 & otherwise \end{cases} \tag{1}$$

The number 0.2 is arbitrary, but it means that the face-hair variation is 5 times less important than variations involving the background of the image.

Let F be the function that measures the variation of face-frame between two images (I_1 and I_2) of same size, with height h and width w. The classifications of the pixels of I_1 are $p_{i,j}$, and the classifications of the pixels of I_2 are $q_{i,j}$.

$$F(I_1, I_2) = \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} f(p_{i,j}, q_{i,j})}{h * w} \tag{2}$$

This function yields a percentage of the images that is classified differently, giving more importance to background variations. If the face-frame does not

vary between two images I_1 and I_2, then $F(I_1, I_2) = 0$. If there is a change in at least one pixel, then $0 < F(I_1, I_2) \leq 1$.

2.3. Face frame correction algorithm

To try to obtain results with the least variation of face frame as possible, we move in the surroundings of the latent space of the original face and minimize the function defined in 2.2.

The whole algorithm is very similar to the projection algorithm designed for StyleGAN2 [5]. The correction takes the target image and a latent code (usually output of projection or result of moving through latent direction) as its input. First, it calculates the standard deviation of 10000 random latent codes that create realistic images, which allows us to apply noise to the original latent code in a way that we can be certain it will yield a realistic image. The $f(targetimage, G(initiallatentcode))$ is also calculated and stored in this first step. The noise strength in each iteration is the result of applying the following formula:

$$strength(i) = l_{std} * noise_0 * (\frac{1 - \frac{i}{10000}}{nrl})^2 \tag{3}$$

Being i the current iteration number, l_{std} the standard deviation of the latent codes previously mentioned, $noise_0$ a constant representing the initial noise factor, and nrl a constant representing the noise ramp length. In these experiments, the values are $noise_0 = 0.005$ and $nrl = 0.75$, which follow what StyleGAN2 uses [5].

Throughout every of the n iterations, the algorithm introduces a Gaussian noise multiplied by $strength(i)$ to the latent code. If this new latent code generates an image with less face-frame variation than the previous latent code, then this new one is stored and used as the current latent code. If that is not the case, the original is preserved. In Figure 3, we can see an example of a target face being projected into the latent space as it outputs the neural network and another one after applying the correction algorithm.

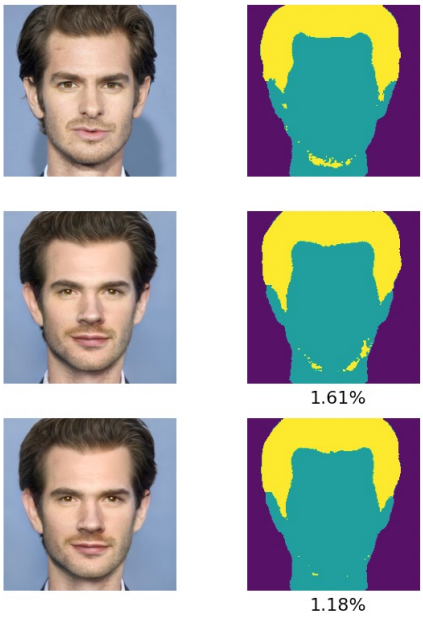


Figure 3: A target face image (top left), its output from the neural network (top right) and its output after running it through the correction algorithm, with its face frame difference.

Having set fixed values of $noise_0 = 0.005$ and $nrl = 0.75$ taken from StyleGAN2 [5], the optimal value of n is studied having in mind the reduction of the face-frame variation of the target image and the projected one as much as possible.

3. Results

3.1. Projection to latent space

Table 1 shows the face frame difference that the image to latent space projection achieved with each image. The maximum difference is at 3.192% while the minimum at 0.747%. The mean difference between these images is of 1.812%. This results suggest that images are a decent starting point when trying to generate similar faces with minimum face frame difference.

Image	Andrew	Daniel	Emma	Jennifer	Kristen	Matt
Difference	1.429%	1.894%	2.531%	1.790%	2.267%	1.452%

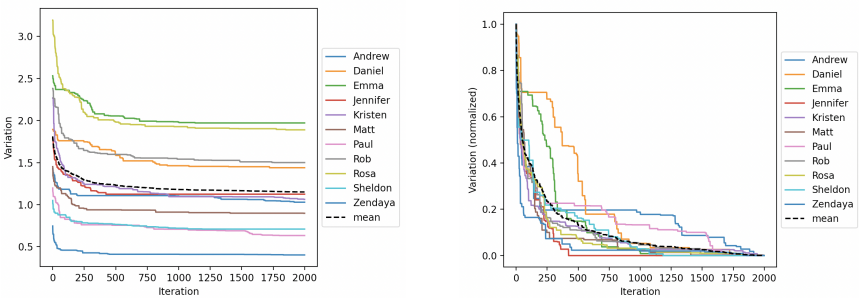
Image	Paul	Rob	Rosa	Sheldon	Zendaya
Difference	1.198%	2.381%	3.192%	1.041%	0.747%

Table 1: Face frame difference results for projecting to latent space

Figure 4: The 11 chosen faces and their respective projection generated by StyleGAN2 to their bottom.

3.2. Correction for projection

Figure 5 illustrates the improvement in the face frame difference through the iterations of the correction. The curves of variation significantly decrease in the first 750 iterations. The last 1250 iterations takes account for the last 10% of the improvement. It is notable that the improvement in the variation is always superior to the 20% of the initial value, and in some of the cases, it goes over the 50%.



(a) Absolute (b) Normalized

Figure 5: Graphs of the face frame difference through iterations

3.3. Latent directions

We also check the face frame difference for the 10 images when transforming those moving across the some of the latent directions known by the community [6]. In Figure 6, we can see the chosen images. They are labeled as 1 to 10 from left to right.

Figure 6: The 10 chosen faces for the latent directions experiments.

When an image moves a great magnitude through a direction, they tend to give unexpected results, deforming the base structure of the face and giving unrealistic images. This is the reason the range measured in each direction changes in each case. For the set of chosen images, we decided the following ranges:

Direction	Age	Gender	Vertical	Horizontal	Eyes-open	Mouth-open	Smile
Range	[-3, 3]	[-3, 3]	[-3, 3]	[-3, 3]	[-10, 10]	[-10, 10]	[-2, 2]

Table 2: Ranges to measure for each direction

Figure 7 shows that the face frame difference is directly proportional to the distance moved in every tested direction. Horizontal alignment has the most impact on the face frame difference when compared to the others. On the other hand, the mouth and eyes opening directions show the least impact.

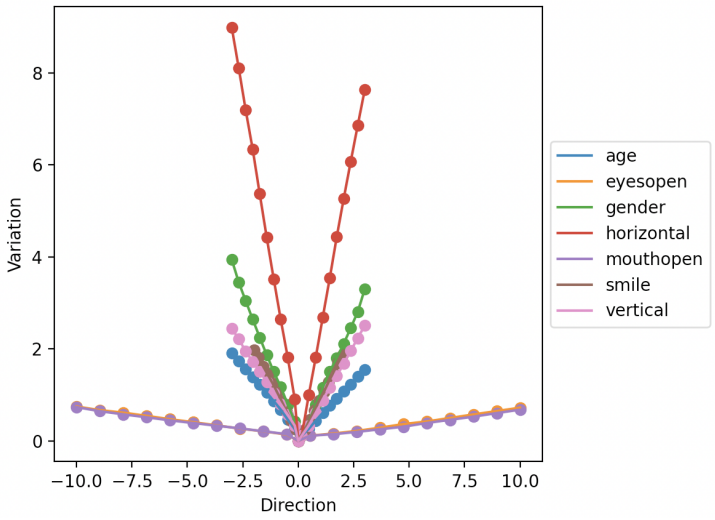


Figure 7: Graph of the mean face-frame variations through the each studied direction

4. Conclusions

The proposed algorithm proved to be useful when trying to project a target image generating a new one which is not only similar to the original one in terms of facial characteristics, but it also keeps most the facial frame as much as possible.

The face frame difference seems to increase with the magnitude of the movement through any of the studied directions in a linear way. However, the rate of such variation is different for each direction. Indicating that there are features that have a bigger influence in the face frame than others according to the neural network model.

The proposed work is a step toward better perceived quality in face editing techniques with StyleGAN and the better understanding of the latent space.

References

[1] S. Want, O. Pascalis, M. Coleman, M. Blades, Recognizing people from the inner or outer parts of their faces: Developmental data concerning 'unfamil-

iar' faces, British Journal of Developmental Psychology 21 (2003) 125 – 135. doi:10.1348/026151003321164663.

[2] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[3] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.

[4] K. Gupta, Face segmentation, https://github.com/kampta/face-seg (2018).

[5] NVlabs, Stylegan 2, https://github.com/NVlabs/stylegan2 (2021).

[6] R. Luxemburg, *StyleGAN2 latent directions*, https://twitter.com/robertluxemburg/status/1207087801344372736 [Accessed: 2022-08-10] (2019).