Mini-Project 4 - Paper Reproducibility Challenge

Mario Francisco Munoz, Wassim Wazzi, Cesar Arnouk
COMP 551 - Applied Machine Learning

Summary of the paper

Blind face restoration is a task that normally relies on the geometric information in a face to then reconstruct that face. The problem is that the information present in a degraded image is noisy and not very clear. This paper proposes a way to use different "geometric priors" from a wide set of faces that are learned by a GAN to perform blind face restoration. It achieves this by adding a spatial transform layer to the model.

Scope of Reproducibility

We attempt to reproduce similarity scores detailed in the paper and then proceed to push the model to its breaking point by adding noise and compression to the input images until there is a radical change in the similarity of the images coming out of the model. We also tested the model on different databases to see if occlusion, lighting, age, pose and other factors influenced the model's ability to recreate faces.

Methodology

We used the same code as the authors. This was particularly easy as the clean, documented code was available on GitHub. However, the training of the model was particularly difficult because of its complex structure, but also because of the sheer amount of data needed to train it. Because of this, we used the pre-trained facial restoration model provided by the authors of the paper for all the experiments. Using image libraries, we added noise and compressed the original images from multiple databases to simulate natural degradation. We selected a sample of 100 images from each dataset and proceeded to feed these examples through our pipeline. We evaluated the performance of the model using the same similarity metrics that the authors detail in the paper. A qualitative analysis of the resulting images was also performed, hypothesizing the different artifacts imagined by the generative model and discussing possible sources of failure given the current model architecture.

1 Model Analysis

1.1 Architecture

The model begins by implementing a degradation removal solution by using a U-net (Ronneberger, Fischer, and Brox 2015) architecture to encode and decode the input image. Additionnally, this U-net extracts latent features as well as multi-resolution spatial features. The latent features discovered by the U-net are then passed through a multi-layer perceptron that presents its output to a spatial feature transform algorithm. This is the code that bridges the U-Net to the GAN at every stage of the GAN's generation. The spatial feature algorithm's job is to generate two affine transform parameters (α and β) by taking the U-net's spatial features and applying convolutions. Additional care is provided to ensure a balance of fidelity and realness by using channel-splitting and concatenating the original GAN features. These parameters will modulate the GAN's features. This complex framework can be observed in Figure 3.

1.2 Loss Functions

The model employs various loss functions (see Equation 5) to calculate a total loss function. The first of these, reconstruction loss, is meant to keep the models prediction $\hat{y}$ close to the original y ground truth image (Equation 1). To achieve this, we add two terms. The first is the L1 reconstruction loss distance that can be calculated with $\hat{y}$ and y directly. The second term is calculated using $\phi(\hat{y}), \phi(y)$ denoting a pretrained VGG-19 network. This can be percieved as a perceptual loss. Equation 2 shows the adversarial loss employed by the GAN. This loss function is meant to incite the GAN to generate natural-looking pictures by looking at the descriminator prediction $D(\hat{y})$. Equation 3 shows a facial component loss from a region of interest (i.e. mouth, left eye). Its first term demonstrating the same discriminative loss described in (Goodfellow et al. 2014). The second term in this equation denotes a feature style loss. Equation 4 presents an identity loss, inspired by (Huang et al. 2017). This loss is defined by comparing the identity

of the ground truth image with the prediction image. This is obtained by using a pre-trained model. The authors use ArcFace (Deng et al. 2019). The total loss is presented in Equation 5 as the sum of all losses.

2 Datasets

Different databases and specific images were used for testing troubleshooting, however the main database used to provide comparable results is the "CelebA" database (Liu et al. 2015) as this one was used by the authors in the paper as well. We used two other databases to examine the difference in results. The first is UTK, a large face dataset with occlusions, variations in orientation and changes of lighting in a large span of age groups (*UTKFace* 2021). The second, an even more varied dataset with images of protests, faces with makeup and varying scale (Yang et al. 2016).

2.1 Noise and Compression

The authors add noise in the training (for data augmentation) and test (for natural image degradation simulation) datasets. We ran a data pipeline using OpenCV (Bradski 2000) that compressed images and added noise just as the authors detail in their paper and then tested other alternatives when the model experienced difficulties when inferring. This is a 3-part approach, starting with a gaussian blur kernel, a downsizing of the image, and finally the adding of white noise. This can be observed in Equation 6. The parameters used in these functions came directly from the paper, aiming to maximize the similarity in the degradation applied to the images during the testing phase. Because of our use of different varied datasets, the resolution in the faces varied greatly depending on the proportion of size that the face occupied in the image.

3 Results

(a) In this figure, we can observe a good reconstruction of a face with glasses. We can also note the lack of symmetry in the frames.

(b) In this figure, we can observe the crude detection of features in an occluded face with a radical pose.

Figure 1: Two results from our testing from different databases

The model achieved very impressive results at times, especially when the faces are not obstructed, and degradation levels are not very high. However, when the input is too blurry, or the ratio of the size of the face compared to the image is small, the faces in the output become deformed, and some features are hallucinated on top of the faces (glasses for example).

3.1 Similarity Metrics

The metrics used in this paper are detailed in equations 7 to 11 in order to evaluate the performance of the model on all datasets. When compared to the original paper, we see that the results on the celebA dataset are not quite similar, this could due to many reasons. Firstly, we sub-sampled the dataset in order to maximize the GPU resources at hand, this might have influenced our findings if the batch was abnormally hard to infer. Another factor that could have influenced our findings was a difference in the noise and down scaling functions. Even though we carefully followed the article's guidelines, there could have been a difference in the Gaussian blur, down scaling or white noise functions. The upper bounds on the noise functions that we used were very high and this section was not thoroughly covered by the authors

	PSNR	SSIM	FID	LPIPS
Input_celebA	14.400068	0.531505	222.854079	0.517061
Input_utk	10.373555	0.459653	159.464849	0.566313
Input_wider	10.404472	0.508763	263.366030	0.603343
Output_celebA	13.727466	0.467375	239.039520	0.537694
Output_utk	9.931122	0.414266	148.868457	0.546444
Output_wider	9.595780	0.413394	313.560935	0.740526
GT_celebA	inf	1.000000	-0.000064	0.000000
GT_utk	inf	1.000000	-0.000086	0.000000
GT_wider	inf	1.000000	-0.000059	0.000000

Figure 2: Metric results we achieved with the 3 Datasets

Methods	LPIPS↓	FID↓	NIQE ↓	Deg.↓	PSNR↑	SSIM↑
GFP-GAN (ours)	**0.3646**	**42.62**	**4.077**	**34.60**	25.08	0.6777
Input	0.4866	143.98	13.440	47.94	25.35	0.6848
GT	0	43.43	4.292	0	∞	1

Figure 3: Original results presented in the paper on the celebA dataset.

in the paper. The color jittering function they used was not described. This is emphasized by the difference in FID results. However, when looking at the relative difference between the PSNR in inputs and outputs, the results are more similar to those detailed in the paper. It is also interesting to note that our output under-performs the paper model in every metric.

3.2 Qualitative Evaluation

We present an array of different results, each labeled according to the important behavior witnessed in the image through Figures 3 until 14. Throughout this evaluation we can identify major factors that impede the proper reconstruction. These failure factors include partial face occlusion (see Figure 11, 13), radical poses different from a frontal, eye-level photograph (see Figure 7,9), improper or unbalanced lighting (see Figure 7, 12, 14), poor resolution (see Figure 5,6), lack of facial symmetry (see Figure 4,7,10) and lack of visible ethnic minorities in the dataset (see Figures 8,9). These factors all influence the clarity of the geometric priors that the network can extract from the original image and obviously gravely impact the reconstruction. It would be interesting to see an implementation of this model that allows the hard coding of certain latent facial features. For example, if I knew that my grandmother didn't wear glasses and had brown eyes, the sex, glasses and eye colour information could be passed on to the model to better the reconstruction of the original image. Mapping these latent codes could also contribute to a better understanding of the model's decisions and open a discussion about explicative artificial intelligence on the subject, as having an unexplained model could potentially limit its applications.

4 Discussion

All in all, the model tends to perform properly, but not in every case. It achieves proper face restoration for images than have gone through a reasonable amount of degradation. However, when we added noise that affected one or more of the aforementioned failure factors too much, the model started to show some inconsistencies (as can be seen in figures 4 to 14). This is homogeneous with the results in the paper, as the authors stated that "when the degradation of real images is severe, the restored facial details by GFPGAN are twisted with artifacts."

5 Appendix

5.1 Equations

- Reconstruction loss

$$L_{reconstruction} = \lambda_{l1}||\hat{y} - y||_1 + \lambda_{per}||\phi(\hat{y}) - \phi(y)||_1 \tag{1}$$

- Adversarial loss

$$L_{adversarial} = -\lambda_{adversarial} E_{\hat{y}} softplus(D(\hat{y})) \tag{2}$$

- Component loss

$$L_{component} = \sum_{ROI} \lambda_{local} E_{\hat{y}ROI}[log(1 - D_{ROI}))] + \lambda_{fs}||Gram(\psi(\hat{y}_{ROI})) - Gram(\psi(\hat{y}_{ROI}))||_1 \tag{3}$$

- Identity preservation loss

$$L_{identitypreservation} = \lambda_{id}||\eta(\hat{y}) - \eta(y)||_1 \tag{4}$$

- Total loss

$$L_{total} = L_{rec} + L_{adv} + L_{comp} + L_{id} \tag{5}$$

- Noise and Compression equation:

$$x = [(y \cdot k_\sigma) \downarrow_r + n_\delta] JPEG_q \tag{6}$$

Values used by the authors:
the gaussian blur kernel was unspecified, so we used a range between 3 and 15

$$\downarrow_r = 1:8, \sigma = 0.2:10, n_\delta = 0:15 \tag{7}$$

- FID - Fréchet Inception Distance (Heusel et al. 2018)

$$FID = ||\mu - \mu_w||_2^2 + tr(\Sigma + \Sigma_w - 2(\Sigma \cdot \Sigma_w)^{1/2}) \tag{8}$$

- NIQE - Natural Image Quality Evaluator (Mittal, Soundararajan, and Bovik 2013)

$$D(v_1, v_2, \Sigma_1, \Sigma_2) = \sqrt{((v_1 - v_2)^T)(\frac{\Sigma_1 + \Sigma_2}{2})^{-1}(v_1 - v_2))} \tag{9}$$

- LPIPS - Learned Perceptual Image Patch Similarity (Zhang et al. 2018)

$$d(x, x_0) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} ||w_l \cdot (\hat{y}^l_{hw} - \hat{y}^l_{0hw})||_2^2 \tag{10}$$

- PSNR - Peak Signal to Noise Ratio

$$PSNR = 20 \cdot log_{10}(\frac{MAXf}{\sqrt{MSE}}) \tag{11}$$

- SSIM - Structural Similarity

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma + c_2)}{(\mu_x^2\mu_y^2 + c_1)(\sigma_x^2\sigma_y^2 + c_2)} \tag{12}$$

5.2 Figures

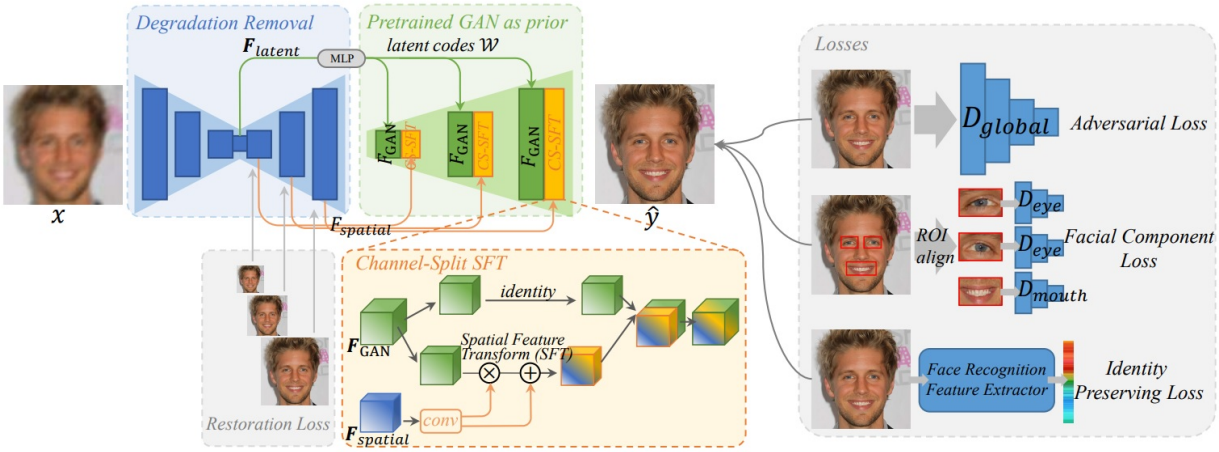

Figure 4: Overview of GFP-GAN framework

Figure 5: In this figure, we can observe the perfect case scenario for our model, a well-lit face without occlusion.

Figure 6: In this figure, we can observe a feature mismatch of a surgeon with two very different eyes and a deformation around the mouth.

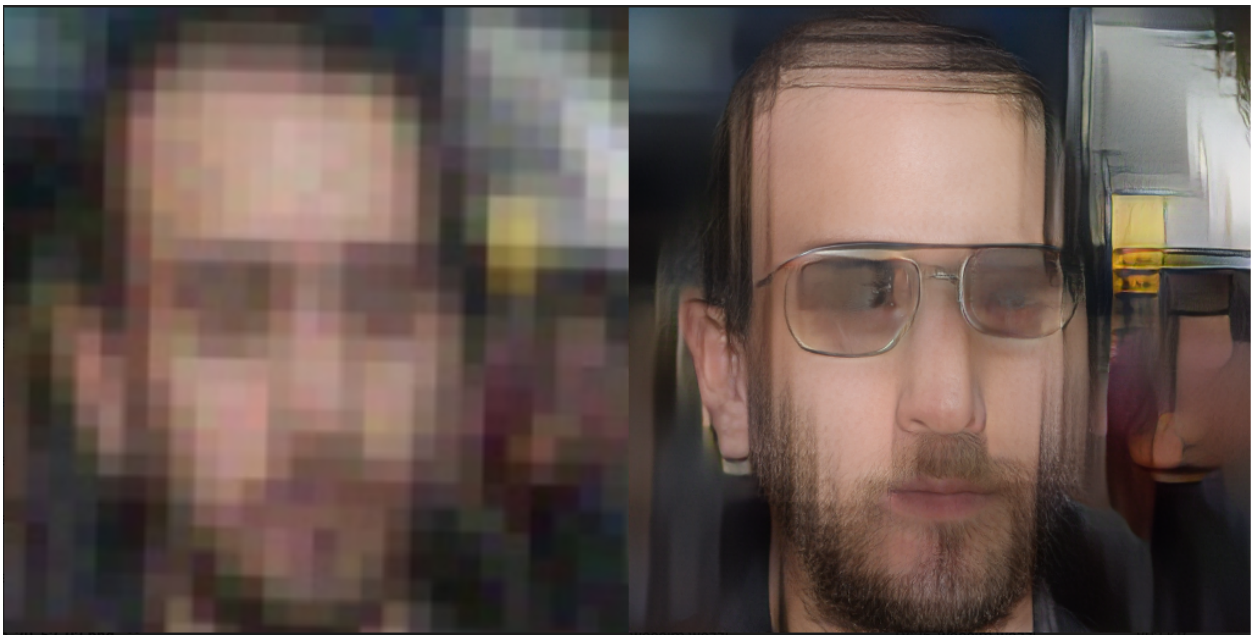

Figure 7: In this figure, we can observe the hallucination of glasses by the GAN model. Face alignment makes pixel distortion look like glasses.

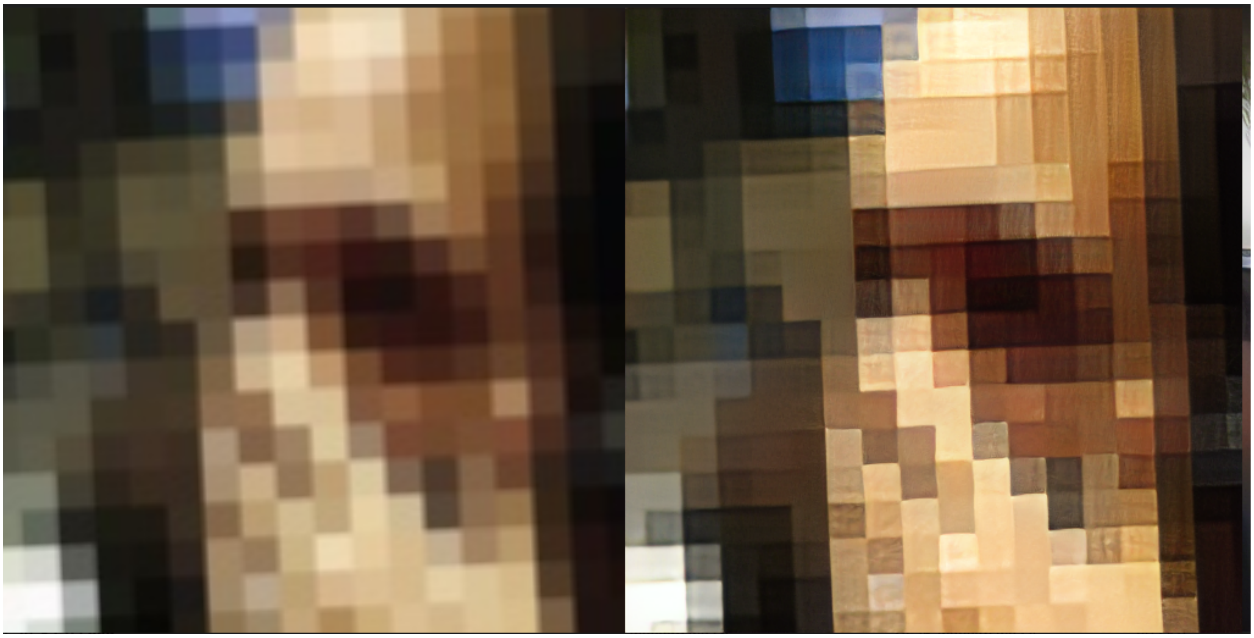

Figure 8: In this figure, we can observe a total failure at an attempt where the model recognized a face, but had extracted little geometric information

Figure 9: In this figure, we can observe another mismatch between features. The left and right eyes are not matched and the glare in the subject's left eye appears to be coming from a camera flash. Could the model be biased? After all,it was trained on a celebrity dataset.

Figure 10: In this figure, we can observe an odd texture around the man's hair, indicating maybe fewer examples of ethnic minorities were represented in the training data as the model struggles to find the right hair texture.

Figure 11: In this figure, we can observe a face with a difficult pose. The model does not perform as well.

Figure 12: In this figure, we can observe another eye feature mismatch with the same flash lighting phenomenon.

Figure 13: In this figure, we can observe the model's failure to reconstruct the glasses on the man's face.

Figure 14: In this figure, we can observe an eye alignment mismatch.

Figure 15: In this figure, we can observe a more difficult reconstruction for the model, given that the face is occluded and the model has no geometric information about hands.

Figure 16: In this figure, we can observe another eye color and alignment mismatch.

References

Bradski, G. (2000). “The OpenCV Library”. In: *Dr. Dobb’s Journal of Software Tools*.

Mittal, Anish, Rajiv Soundararajan, and Alan C Bovik (2013). “Making a ‘Completely Blind’ Image Quality Analyzer”. en. In: p. 4.

Goodfellow, Ian et al. (2014). “Generative Adversarial Nets”. en. In: p. 9.

Liu, Ziwei et al. (Dec. 2015). “Deep Learning Face Attributes in the Wild”. In: *Proceedings of International Conference on Computer Vision (ICCV)*.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (May 2015). “U-Net: Convolutional Networks for Biomedical Image Segmentation”. en. In: *arXiv:1505.04597 [cs]*. arXiv: 1505.04597. URL: http://arxiv.org/abs/1505.04597 (visited on 11/28/2021).

Yang, Shuo et al. (2016). “WIDER FACE: A Face Detection Benchmark”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, Rui et al. (Oct. 2017). “Beyond Face Rotation: Global and Local Perception GAN for Photorealistic and Identity Preserving Frontal View Synthesis”. en. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Venice: IEEE, pp. 2458–2467. ISBN: 978-1-5386-1032-9. DOI: 10.1109/ICCV.2017.267. URL: http://ieeexplore.ieee.org/document/8237529/ (visited on 12/02/2021).

Heusel, Martin et al. (Jan. 2018). “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium”. en. In: *arXiv:1706.08500 [cs, stat]*. arXiv: 1706.08500. URL: http://arxiv.org/abs/1706.08500 (visited on 12/07/2021).

Zhang, Richard et al. (2018). “The Unreasonable Effectiveness of Deep Features as a Perceptual Metric”. en. In: p. 10.

Deng, Jiankang et al. (2019). “ArcFace: Additive Angular Margin Loss for Deep Face Recognition”. en. In: p. 10.

UTKFace (2021). en-US. URL: https://susanqq.github.io/UTKFace/ (visited on 12/13/2021).