GAN Inversion for Occlusion Removal in Face Images

Gonçalo Alexandre Pires de Mendonça Teixeira
Instituto Superior Técnico - Universidade de Lisboa, Portugal
Email: goncalo.m.teixeira@tecnico.ulisboa.pt

***Abstract*—Facial occlusions present a challenging problem in computer vision, affecting various applications such as face recognition and image analysis. This work introduces a novel approach that leverages Generative Adversarial Networks (GAN) inversion techniques to remove occlusions from face images. The proposed method incorporates prior knowledge by utilizing a set of face images without any occlusion as a reference during the reconstruction process. By inverting the latent codes of the occlusion-free reference images, a representative latent space is established. Notably, this approach distinguishes itself from existing methods by incorporating additional information about the person's face through the use of reference images. During the optimization process, each reference image's latent code is associated with a Gaussian function which guide the optimization towards reconstructing the occluded image while considering the unique facial characteristics present in the reference images. By fusing both the occluded image and prior knowledge, the proposed method aims to produce more accurate and realistic results. Experimental results demonstrate the efficacy of the proposed method for occlusion removal, which is further validated through the utilization of the 2D-FAN algorithm for facial keypoints localization.**

***Index Terms*—Face reconstruction, Occlusions, Generative adversarial networks, GAN inversion, Latent space.**

I. Introduction

Facial occlusions present a pervasive and challenging problem in computer vision [1], impacting a wide range of critical applications such as face recognition, image analysis, biometric systems, and human-computer interaction. These occlusions, resulting from objects or obstructions partially covering the face, introduce uncertainties and errors in facial analysis algorithms, posing significant challenges to the reliability and performance of these systems. For instance, in surveillance systems, facial occlusions might hinder the identification of individuals, leading to potential security concerns and missed opportunities to prevent criminal activities.

Traditional methods for occlusion removal rely on heuristic approaches and handcrafted features, but they often fall short in handling complex occlusions effectively. These approaches may lead to incomplete or distorted facial reconstructions, limiting their practical applicability in real-world scenarios. Recent advancements in deep learning and Generative Adversarial Networks (GAN) [2] have shown immense promise in various image-related tasks, raising hopes that they can tackle the formidable challenge of facial occlusion removal. However, despite the steady progress in GAN-based inversion methods [3], accurately reconstructing occluded face images remains a persistent challenge. Existing GAN inversion methods excel at inverting unoccluded face images into their latent space, but when it comes to occluded images, they tend to produce suboptimal results with blurred regions and compromised facial features.

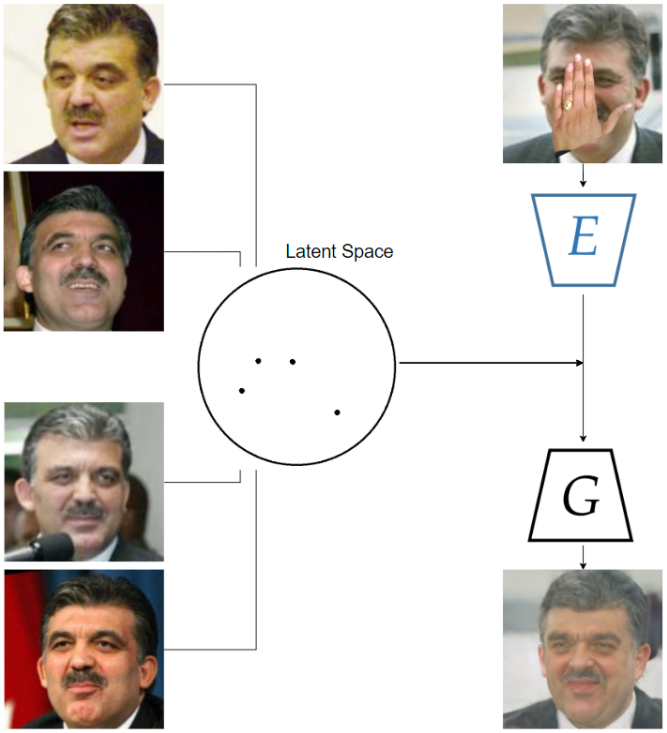

Fig. 1. GAN Inversion model architecture using "clean" images as prior knowledge during optimization.

This work aims to address the shortcomings of current approaches by introducing a novel method that leverages GAN inversion techniques for precise and realistic facial occlusion removal. By capitalizing on the representational power of GANs and their ability to learn rich latent spaces encoding essential facial characteristics [4], we strive to achieve accurate and visually plausible reconstructions. The proposed method incorporates prior knowledge about unoccluded facial appearance by utilizing a set of reference images without any occlusions. This integration of reference images empowers the optimization process, guiding it towards preserving the individual's unique facial features while removing the occlusions.

We summarize our contributions as follows:

- Investigate the challenges posed by facial occlusions in

computer vision and highlight the limitations of existing methods in handling complex occlusions.
- Explore the capabilities of GANs and their potential to reconstruct facial images while preserving essential facial characteristics.
- Develop a novel approach that incorporates prior knowledge using reference images without occlusions, and investigate its impact on the accuracy and realism of occlusion removal.
- Evaluate the performance of the proposed method and validate the effectiveness of the proposed method for downstream facial analysis tasks, such as facial keypoints localization, to demonstrate its potential applicability in real-world scenarios.

II. Background

This chapter contains a brief explanation of the background on image inpainting using GANs. It starts with a general description of GANs, how its latent space works and an explanation of GAN Inversion techniques.

A. *Generative Adversarial Networks - GANs*

GANs [2] have emerged as a pivotal innovation in the realm of deep learning, facilitating the generation of synthetic data that aligns with the distribution of training data. GANs stand apart from other discriminative models by focusing on capturing the underlying data distribution, thereby enabling the creation of new instances that plausibly adhere to this distribution. This paradigm shift has profound implications for data synthesis and generation.

*1) **Architecture**:* The architecture of a GAN comprises two key components: a generator (G) and a discriminator (D). These components engage in a dynamic adversarial interplay, wherein the generator aims to convert a simple random vector sampled from a distribution into a more complex output that follows real data instances (in our case face images), while the discriminator aims to differentiate between real and generated data. The adversarial process encourages both components to refine their capabilities iteratively, ultimately leading to the generator producing data that is indistinguishable from real data samples.

*2) **Training the model**:* The training dynamics of GANs involve a zero-sum game, where the generator's pursuit of creating realistic data instances challenges the discriminator's capacity to distinguish real from generated samples. This adversarial process is mathematically encapsulated in the objective function of Equation 1, where $L(D, G)$ represents the adversarial loss to be minimized by the generator and maximized by the discriminator, G stands for the generator, D for the discriminator, x is a sample from the data distribution p_{data}, and z is a noise sample from a noise distribution p_z. The generator's output, along with real data instances, is assessed by the discriminator, thereby driving both networks toward their respective optimization objectives.

$$\min_G \max_D \; L(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{1}$$

This adversarial training ends when the generator successfully generates data that is convincingly realistic. The resultant equilibrium implies that the generator has become adept at generating data that closely matches the distribution of the training data, effectively indistinguishable from authentic instances.

*3) **StyleGan**:* All experiments conducted in this work utilize the StyleGAN architecture, which is an advanced generative model designed for creating high-quality images with fine-tuned control over their appearance. Introduced in the work by Karras et al. [5], StyleGAN is known for its ability to generate lifelike human faces with exceptional realism. At its core, StyleGAN employs a mapping network to transform a random noise vector into a "style" vector, which serves as a control mechanism for image attributes such as pose, age, and more. The synthesis network, another critical component, utilizes this style vector to progressively generate images, allowing for detailed and customized outputs.

One of the remarkable strengths of StyleGAN lies in its capacity to capture intricate facial details, including realistic skin texture, fine lines, and subtle expressions. Moreover, it offers control over various attributes such as age, gender, and ethnicity, making it capable of producing a diverse range of faces with remarkable fidelity.

B. *GAN's Latent Space*

The earlier GAN works predominantly concentrated on image manipulation, facilitated by the intrinsic capabilities of GAN latent space arithmetic. At the core of GAN operations lies the generator's profound ability to map points within the latent space to distinct characteristics within the generated images. This latent space, an N-dimensional domain learned by the GAN during training, is unique for each trained GAN model.

Notably, the latent space can be leveraged to effect specific changes within images by modifying the latent codes in semantically meaningful directions. However, it is essential to emphasize that this form of manipulation is directly applicable only within the context of images synthesized by the GAN.

Within the latent space, binary semantics, such as age distinctions (young and old), can be encapsulated by hyperplanes, functioning as separation boundaries, Both Shen *et al.* [6] and Alec Radford *et al.* [7] highlighted this concept, showing how a boundary could be employed to manipulate the orientation of faces within images. They based this hyperplane on averaged samples of faces looking left versus those facing right. They also demonstrated the phenomenon of vector arithmetic within the latent space, resulting in the emergence of multiple distinct visual concepts within the generated images.

C. *GAN Inversion*

The primary goal of GAN inversion is to extract the latent code for every given image, so that when it is sent through the generator, it generates an image as close as possible to the input one. In the context of inpainting, GAN inversion's objective revolves around extracting latent codes that, when processed by the generator, yield images closely resembling

the original unaltered counterparts. Existing GAN Inversion approaches usually fall into three types.

Beginning with the optimization-based method [8], the latent code z is optimized iteratively to reconstruct an image. However, the non-convex nature of this optimization process makes this method computationally intensive and time-consuming.

On the other hand, the learning-based technique [9] focuses on training an encoder to map images to their latent codes. The encoder's goal is to deduce a latent code that, when passed through the generator, yields an image that faithfully reconstructs the input image. This method is significantly more efficient than the optimization-based as the latent code can be directly inferred by passing the image through the encoder during inference. Nonetheless, it is essential to note that a single inference may not always identify the closest latent code due to the nuanced complexities of the latent space.

Considering the characteristics of both methods, we opted for a hybrid-based approach, where we try to combine the strengths of both methods. In this methodology, an encoder searches for a latent code during inference, initiating the optimization-based approach to refine the latent code and minimize discrepancies between the initial and desired latent codes. This hybridization capitalizes on the efficiency of the encoder and the precision of optimization, offering a balanced trade-off between computational cost and accuracy.

III. Methodologies

This chapter gives a brief explanation of the strategies adopted in this work.

A. Model Components

1) ***Baseline****:* The baseline adopted as a starting point in this work was developed by Zhu *et al.* [10] and is based in an in-domain GAN inversion technique. The rationale behind this baseline selection is its proven success in generating high-quality inversion results for images without occlusions.

This baseline employs a hybrid GAN inversion approach, leveraging the StyleGAN model [5] trained in the FFHQ dataset [5]. A key feature is the utilization of a latent space $\mathcal{W}$, derived from mapping the initial latent space $\mathcal{Z}$ through a Multi-Layer Perceptron (MLP) [11], facilitating enhanced disentanglement of semantic features. For simplicity, we will continue to denote the latent code as z in the following sections.

2) ***Gaussian Functions****:* Diverging from conventional GAN Inversion techniques, our proposed approach capitalizes on the integration of prior knowledge by incorporating a set of "clean" facial images from the same occluded user as reference images. These reference images play a pivotal role in integrating prior information into the occlusion removal process. Our solution addresses the occlusion removal problem from a fresh perspective, developing a novel strategy known as an "energy function" centered around the latent codes of reference images as illustrated in Figure 2.

The methodology starts by inverting all the reference images into the latent space of the underlying GAN, generating a comprehensive list of points that collectively encapsulate the user's facial characteristics within the GAN's latent domain. Gaussian functions are then employed to quantify the relationship between the latent code under optimization and the latent codes of reference images, allowing the algorithm to make informed decisions during the optimization process.

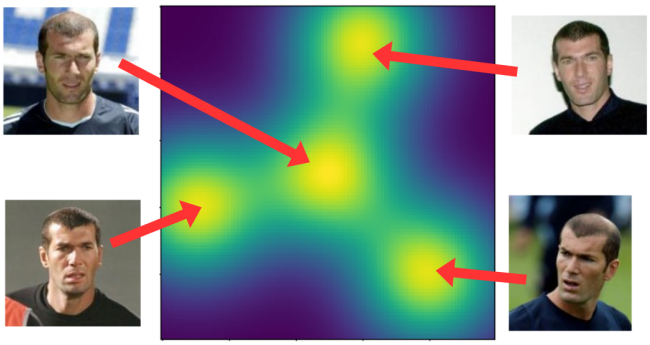

Fig. 2. An illustration of the Gaussian functions effect in the GAN's latent space.

Gaussian functions are centered at each reference image's latent code, and their influence on the latent code under optimization is calculated at each optimization step. A fixed σ value, where all components are equal ($\sigma_{X_1} = \sigma_{X_2} = ... = \sigma_{X_N}$), is utilized along with the norm between the latent code under optimization and each reference image's latent code. The cumulative influence of all Gaussian functions is then summed up.

3) ***Discriminator Loss****:* Traditionally, the discriminator model plays a pivotal role solely during the training of a GAN, primarily aiming to differentiate between real and generated samples. A high discriminator score indicates a more realistic face image. In our approach, we introduce an additional loss term based on the discriminator's output to enhance the realism of the occlusion-free generated image. The discriminator serves as a critical judge, providing feedback to guide the occlusion removal process toward producing images that are not only occlusion-free but also visually convincing and indistinguishable from genuine facial images.

4) ***Final Approach****:* With the inclusion of the discriminator-based and Gaussian functions loss terms into the baseline optimization process, the objective function captures the multifaceted approach employed to comprehensively address occlusion removal.

With the inclusion of this discriminator-based loss term, the objective function, as represented in Equation 2, encapsulates the multifaceted approach employed to tackle occlusion removal comprehensively. This function combines the perceptual loss, reconstruction loss, Gaussian loss, and the discriminator loss, harmoniously working towards the goal of producing occlusion-free images that are not only faithful to the individual's facial characteristics but also exhibit a high degree of realism.

$$z^{inv} = \underset{z}{arg\,min}\ \{\lambda_1 \mathcal{L}_{reconstruction}(z) + \lambda_2 \mathcal{L}_{perceptual}(z) + \\ + \lambda_3 \mathcal{L}_{gaussian}(z) + \lambda_4 \mathcal{L}_{discriminator}(z)\} \tag{2}$$

Locking at Equation 2, all lambdas correspond to the weights given to each loss terms and z to the latent code being optimized. All loss terms are enumerated as follows, where G is the GAN's generator, D the GAN's discriminator and x the input image:

1) **Reconstruction Loss Term:** Minimizes the mismatches between the input image (x) and the generated one $G(z)$, obtained from the current inverted code.

$$\|x - G(z)\|_2 \tag{3}$$

2) **Perceptual Loss Term [12]:** Focuses on preserving high-level semantic information and vital visual features essential for human perception. It ensures semantic consistency using a pretrained VGG model [13] (F) by minimizing the mismatch of the semantic attributes between the input image and its generated counterpart.

$$\|F(x) - F(G(z))\|_2 \tag{4}$$

3) **Gaussian Loss Term:** Maximizes the cumulative influence of all Gaussian functions, each refering to a reference image latent code. z_{ref} is the set of known latent codes, z is the latent code being optimized and σ is the Gaussian sigma.

$$\left(-\sum_{i=1}^{N} \exp\left(-\frac{\|z_{ref}[i] - z\|}{2 \cdot \sigma^2}\right)\right) \tag{5}$$

4) **Discriminator Loss Term:** Maximizes its score in order to increase the realism of the generated image.

$$(-D(G(z)) \tag{6}$$

This multifaceted loss function encapsulates the heart of our occlusion removal methodology, driving the algorithm to optimize the latent code with a keen eye on both fidelity to the reference images and the attainment of a photorealistic final output.

*5) **Proportionality between the occlusion size and the loss terms**:* In pursuit of a more sophisticated and context-aware approach, we introduced a method that dynamically adjusts the weights assigned to the reconstruction loss and Gaussian loss terms. This adaptability depends on the percentage of occluded face pixels. The underlying principle is simple yet powerful: the algorithm's weight assignment should reflect the extent to which the facial characteristics are obscured by occlusion, thereby optimizing the occlusion removal process.

When a substantial portion of the face is occluded, the Gaussian loss should carry a higher weight than the reconstruction loss, and vice versa when the occlusion is less severe. If the algorithm does not have much information about the facial characteristics of the individual in the image being reconstructed it should be looking for it in the reference images domain, giving more importance to the gaussian loss term in comparison to the reconstruction one. On the other hand if the facial characteristics are much present on the image being reconstructed, the algorithm does not need to search it in the reference images domain, giving more importance to the reconstruction loss term in comparison to the gaussian one.

The weighting system was implemented with a sigmoid function receiving the percentage of occluded pixels as shown in Equation 7, where k and x_0 are the scale and shift parameters and x is the occlusion percentage obtained.

$$\lambda_3 = 2 \cdot \frac{1}{1 + \exp(-k \cdot (x - x_0))} \tag{7}$$

As you can see, the calculated weight for the Gaussian loss is set to be equal to the value derived from this expression, while the reconstruction loss term is fixed to 1.

By aligning the weighting scheme with the degree of occlusion, this approach ensures that the algorithm allocates its resources appropriately, striking the right balance between preserving existing features and seeking information from reference images.

B. Hyperparameter Tuning

Looking at the optimization process architecture 2, the hyperparameters that need tuning are identified as the constants that influence the model's performance, as following listed regarding the proportional weighting system for the reconstruction/Gaussian loss terms, the discriminator loss term and the Gaussians sigma:

- **Scale k:** The scale parameter for the sigmoid function 7 controls how steep the transition between the reconstruction loss and the Gaussians loss terms will be. A higher value will result in a sharper transition, while a lower value will create a more gradual transition.
- **Shift x_0:** The shift parameter for the sigmoid function 7 determines the midpoint of the transition. It is the occlusion size at which the weighting function output is approximately 0.5.
- **Discriminator loss weight λ_4:** This weight represents the importance given to the discriminator score in the combined loss 2.
- **Gaussian sigma σ:** The Gaussian functions graph have a symmetric "bell curve" shape and the sigma parameter (standard deviation) controls the width of the "bell". A higher sigma value result in a wider "bell".

The hyperparameter values space is described in table I.

TABLE I. Hyperparameter Values Space.

Hyperparameter	Space of Values
scale	0.25
shift	[8, 10, 11, 13]
Discriminator loss weight	[0.1, 0.5, 1.0, 2.0]
Gaussian sigma	[5.0, 7.0, 10.0, 15.0]

The scale and shift values are chosen based on the percentage of occluded pixels for small, medium and large occlusions and its values are explained in section V. The discriminator loss weight and Gaussian sigma space of values are set through

preliminary experiments. For the discriminator loss weight, when given high values (greater than 3.0) the resulting image started to be completely different from the target one. Since the primary goal is to remove occlusions while maintaining the realism of the generated content, the model should ensure that the discriminator's feedback is still valuable but not overly dominant. Realism is important, but not at the expense of completely altering the target image.

Between the most common hyperparameter search methods, we choosed the grid search approach, since we had already a ballpark range of known hyperparameter values that perform well.

IV. Experimental Setup

In this chapter, we detail the hardware and software environment used to run the experiments. We describe the dataset used, including its source of the data and any preprocessing steps applied and we also explain the metrics chosen to evaluate the effectiveness of our method and the final hyperparameters used.

A. Network Settings and Computational Environment

The experimental framework was implemented using Pytorch, an Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz and two NVIDIA GeForce GTX 1080 Ti. All models were runned with a learning rate of $\eta = 0.01$, using the Adam optimizer [14]. The batch size was set to 1.

B. Data Manipulation

During the execution of this work some issues needed to be corrected both in data and training.

1) ***Dataset****:* For our research, where both GAN and encoder models are pre-trained, we required a dataset comprising a substantial number of labeled face images organized by individuals. To meet this requirement, we selected the Labeled Faces in the Wild (LFW) dataset [15], which consists of 1680 cases of individuals, each with two or more images, all of which are standardized to a resolution of 250x250 pixels. We inverted all images to the GAN's latent space and due to weak inversion results, the 1680 cases are decreased to 1474, which are split into a validation set (117 cases) which is used to tune the hyperparameters and a test set (1357 cases) which is used to do all the experiments of section V.

2) ***Pre-Processing****:* The initial step in pre-processing involves aligning all faces within the images. We accomplished this using the alignment algorithm from the *Stylegan-Encoder*, which ensures that all faces are centered through facial landmarks localization.

Following alignment, all images must respect the characteristics of the pre-trained GAN model. This involved resizing all images to a resolution of 256x256 pixels and shifting their pixel range to [-1, 1].

Fig. 3. Occlusion generation example.

3) ***Occlusion Generator****:* In our research, we required not only a dataset with multiple images of the same individual but also pairs of occluded images along with their respective binary masks representing the occlusions. To achieve this, we leveraged the work presented in [16] to create synthetic face occlusion examples of high quality. To maintain a sense of realism, we exclusively utilized natural occlusions, specifically occlusions caused by hands [17], like illustrated in Figure 3. Notably, we opted to deactivate certain features in the face occlusion generation process, such as Color transfer (which transfers facial color to the hands) and rotation around the center (which aligns fingers towards the face) due to their significant time-consuming nature.

For each sample within the LFW dataset's 1474 examples, we randomly selected one image to be occluded. This occlusion was applied in three different ways, each involving the same hand occlusion but varying in terms of size and position. Consequently, each sample consisted of one image occluded in multiple ways, reflecting different occlusion sizes and positions.

C. Evaluation Metrics

The performance of the different approaches and hyperparameter configurations were evaluated through four different metrics:

Peak Signal-to-Noise Ratio (PSNR): This widely used metric quantifies the level of distortion or noise present in an image compared to a reference or original image. It is a fundamental measure for evaluating image reconstruction or restoration quality.

Structural Similarity Index (SSIM): SSIM is another well-established metric that assesses the structural similarity between two images. It considers various aspects, including luminance, contrast, and structural information, to provide a score that reflects the perceived quality of an image.

Learned Perceptual Image Patch Similarity (LPIPS): LPIPS [12], a more recent metric, leverages deep learning techniques to evaluate the perceptual similarity between two images. It is specifically designed to capture the visual quality as perceived by humans. However, we observed limitations in LPIPS when assessing reconstruction quality, particularly in cases where only a portion of the image, such as an occluded region, was poorly reconstructed.

TABLE II. Hyperparameter tuning results.

Baseline + Discriminator		Baseline + Gaussians			Baseline + Gaussians + Discriminator		
Discriminator Weight	LPIPS ↓	Shift	Sigma	LPIPS ↓	Shift	Sigma	LPIPS ↓
2.0	0.294	8	15	0.258	8	15	0.258
1.0	0.253	10	15	0.254	10	15	0.254
0.5	0.237	11	15	0.252	11	15	0.252
0.1	**0.226**	**13**	**15**	**0.246**	**13**	**15**	**0.247**

To address this limitation, we introduced a novel metric, which we refer to as **Local LPIPS**. This metric evaluates the LPIPS score exclusively within the reconstructed region of the image. By doing so, we focus on assessing the quality of the reconstructed portion, providing a more accurate reflection of the reconstruction's quality.

Additionally, we incorporated another metric, **2D-FAN** (Facial Alignment Network) [18], into our evaluation. This algorithm is designed to localize facial landmarks in images. It allows us to compare the positions of facial keypoints between the ground truth and the reconstructed image by computing the average difference between both images' keypoints. Like LPIPS and Local LPIPS, a smaller score indicates better performance in preserving facial element positions.

By combining these metrics, we aimed to ensure a comprehensive evaluation of our model's performance. PSNR captures pixel-level similarity, SSIM considers structural details, LPIPS and Local LPIPS assess perceptual quality, and 2D-FAN evaluates the preservation of facial element positions.

D. Hyperparameter Tuned values

We want to compare the reconstruction results with 5 different approaches:

1) Encoder [enc]
2) Encoder + Optimization [baseline]
3) Baseline + Discriminator [disc]
4) Baseline + Gaussian [prior]
5) Baseline + Discriminator + Gaussian [prior_disc] (Expression 2)

This way we tuned the hyperparameters in a progressive way, using a validation set of 117 images (39 small, 37 medium and 41 large) containing all available number of reference images. With the fixed parameters from the baseline (established by the authors), we tuned the discriminator loss weight in the approach number 3. With both baseline parameters and the discriminator weight fixed we tuned the Gaussian constants in the approach number 4. Finally we fixed the discriminator loss weight and tune the Gaussian constants in the final approach (number 5). Although this approach is sub-optimal, we decided to tune the hyperparameters in this way, because of the high computational cost of inverting an image and the high amount of possible hyperparameter combinations. Together with this method we also pre established the scale as 0.25, resuming the number of combinations by $4 + 4 \cdot 4 + 4 \cdot 4 = 36$.

Looking at the average of percentage of occluded pixels for small, medium and large occlusions, and fixing the shift to the average of the medium category, we choose a scale of 0.25 since this gives weights of 0.95 and 0.12 to large and small occlusions.

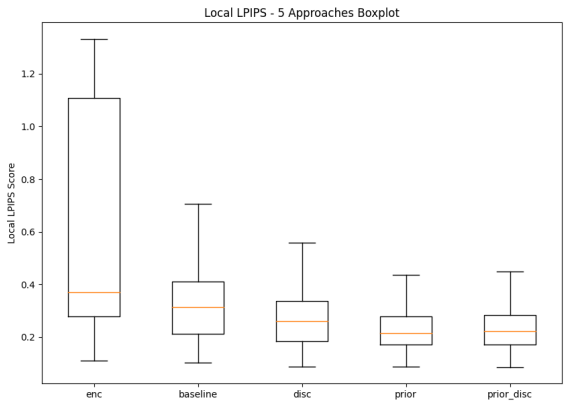

Fig. 4. Local LPIPS boxplot for each approach.

For the discriminator weight, shift and Gaussian sigma the tuned values are 0.1, 13 (which corresponds to the average occlusion size for the "medium" category), and 15 respectively as table II shows. In this evaluation we gave more importance to the LPIPS metric since that was the metric that best correspond to our visual inspection. The LPIPS score problem explained in section IV-C enable us to compare different hyperparameter values inside the same approach, but not to compare hyperparameters along different approaches.

V. Experimental Results

This chapter presents a description and discussion of the experimental results performed during this work.

A. Approaches Comparison

With all hyperparameters tuned, the decision on the method selection was made through three different ways. First we analyzed the local LPIPS and the 2D-FAN metrics and then we compared these measurements with our visual inspection on the reconstructed images.

We can visualize the local LPIPS scores comparison when looking at Figure 4. It is noticeable that both approaches that use our Gaussian methodology have better results. However it is very difficult to distinguish which one is the best approach due to the very similar performances. The same happens

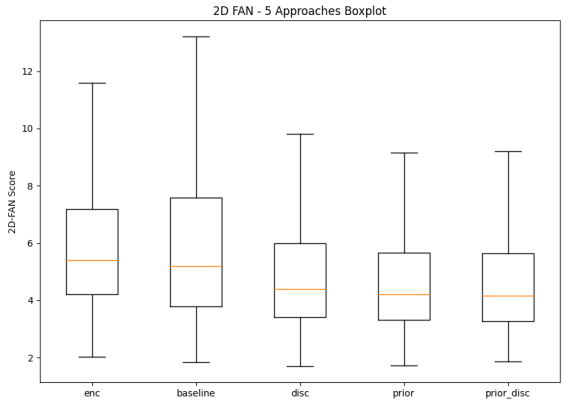

Fig. 5. 2D-FAN boxplot for each approach.

with the 2D-FAN algorithm as Figure 5 shows. However when looking at the mean scores for both metrics as table III indicates, the 'prior' approach has better results in both metrics.

TABLE III. Local LPIPS and 2D-FAN average values for the 5 approaches.

Metric	enc	baseline	disc	prior	prior_disc
Local LPIPS	0.586	0.327	0.273	**0.237**	0.240
2D-FAN (pixels)	7.786	7.980	6.630	**6.262**	6.412

Along with the local LPIPS metric, we chose between the final three approaches through visual inspection of the reconstructed images. Looking at Figure 7, the 'disc' approach is easily excluded as a candidate for the final approach due to the weak reconstruction quality around the occluded portion of the images. Despite the similar results between the other two approaches, we decided to use the 'prior' approach. When looking closely to the first column of 7 is possible to see that the fourth image is brighter (not matching the original) and the person's right eye (occluded one) looks with the wrong dimensions when comparing to the left one.

Fig. 6. Average of local LPIPS score along the amount of the number of reference images. It is noticeable that the reconstruction quality of the image is improving with the increase of the number of reference images used.

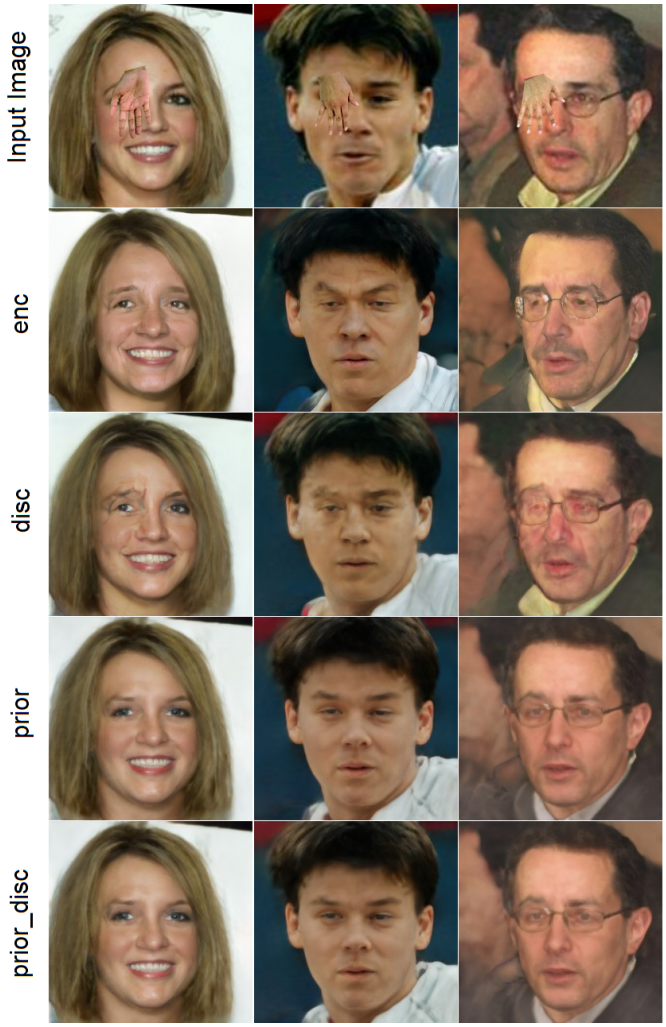

Fig. 7. Qualitative comparison of different reconstruction methods.

B. Analysis on Number of images

As described in section III, our method relies on "clean" face images of the user we want to reconstruct. Obviously there is a trade off between the number of images and the reconstruction quality. To better analyse such trade-off, we evaluate our method by varying the number of reference images used in the reconstruction process of 8 different individuals (with similar number of reference images available).

It is evident from the graph (Fig 6) that, on average, as the number of reference images used in the reconstruction process increases, the local LPIPS score (representing reconstruction quality) tends to improve, indicating better results. This observation aligns with the intuitive expectation that having more reference images allows the model to better understand the user's facial characteristics and produce more accurate reconstructions.

This analysis provides valuable insights into the practical considerations of our approach, especially when dealing with scenarios where the number of available reference images may be limited. It highlights the potential benefits of having a larger dataset of reference images for improved reconstruction quality.

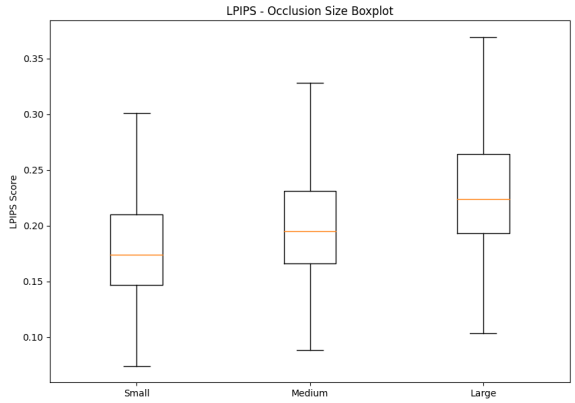

Fig. 8. LPIPS boxplot for each occlusion size category

C. Analysis on Occlusion size/position

As an occlusion removal algorithm, it is important to evaluate how the size and position of the occlusion can affect our model efficiency.

Starting with the occlusion size, the results are pretty straight forward. As expected the model performance decreases with the increase of the occlusion size as Figure 8 shows. For many reference images that we have the best information available comes always from the image being reconstructed, because it is the one we want to match. With the increase of the occlusion size that information is lost, making the model search for that same information on the reference images.

On the other hand Figure 9 contains an histogram that represents how the occlusion position can affect the reconstruction performance of our model. It is noticeable that as the percentage of occlusion within a facial structure increases, there is a discrepancy in the quality of reconstruction between the chin and mouth region compared to other facial structures. This phenomenon is primarily due to the extent of coverage and its impact on the overall appearance.

When an occlusion covers a substantial 80% of the chin or mouth, it essentially masks a significant portion of the face, in contrast with the other facial features like the eyes.

Looking at the smaller occlusions it is noticeable that the left eye is the most complicated facial feature to reconstruct.

Sometimes our model has difficulties in accurately aligning the eyes, resulting in reconstructions where they do not converge as expected. This misalignment, often appearing as a form of strabismus, can be attributed to the combination of loss terms chosen, as there is not a loss term that considers the appearance between both eyes during the reconstruction process.

VI. Conclusion

In this work, we have explored and developed a novel approach for the challenging task of facial occlusion removal using Generative Adversarial Networks (GANs). Our primary objective was to recover the facial features and details obscured by occlusions while maintaining the realism and natural appearance of the reconstructed images.

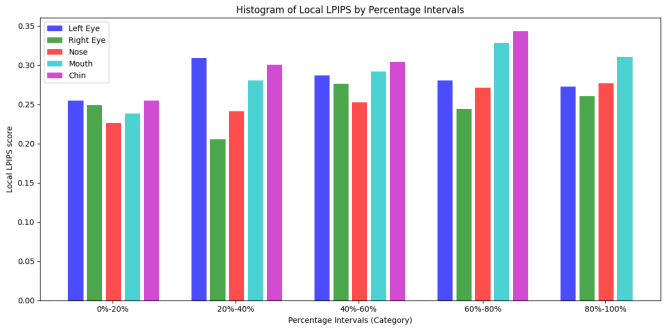

Fig. 9. Histogram of the influence of occlusion positions

One of the key innovations in our approach was the introduction of Gaussian functions as an "energy function" centered around the latent codes of reference images. By quantifying the relationship between the latent code under optimization and the latent codes of reference images, we enabled the algorithm to make informed decisions during the optimization process, having other information sources than the image being reconstructed. This dynamic approach addressed the challenge of occlusion removal in a more context-aware manner.

Our comprehensive evaluation, demonstrated the effectiveness of our approach, as we achieved good results in terms of both technical accuracy and perceptual quality. Moreover, we conducted an analysis of the impact of the number of reference images and occlusion size on the reconstruction quality, providing valuable insights into the trade-offs involved in these parameters.

In conclusion, our work presents a robust and effective solution for facial occlusion removal, with promising applications in various domains. The combination of GAN-based generative modeling and context-aware energy functions opens up new avenues for further research and development in the field of facial image processing and manipulation.

References

[1] Hazım Kemal Ekenel and Rainer Stiefelhagen, "Why is facial occlusion a challenging problem?," in *Advances in Biometrics*, Massimo Tistarelli and Mark S. Nixon, Eds., Berlin, Heidelberg, 2009, pp. 299–308, Springer Berlin Heidelberg.

[2] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial networks," 2014.

[3] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang, "Gan inversion: A survey," 2022.

[4] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou, "Interpreting the latent space of gans for semantic face editing," 2020.

[5] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," 2019.

[6] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou, "Interpreting the latent space of GANs for semantic face editing," .

[7] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," .

[8] Raymond A. Yeh, Chen Chen, Teck Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do, "Semantic image inpainting with deep generative models," .

[9] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas, "Prior guided GAN based semantic inpainting," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 13693–13702, IEEE.

[10] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou, "In-domain GAN inversion for real image editing," .

[11] Simon Haykin, *Neural networks: a comprehensive foundation*, Prentice Hall PTR, 1994.

[12] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018.

[13] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015.

[14] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," 2017.

[15] Gary Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," *Tech. rep.*, 10 2008.

[16] Kenny T. R. Voo, Liming Jiang, and Chen Change Loy, "Delving into high-quality synthetic face occlusion segmentation datasets," 2022.

[17] Mahmoud Afifi, "11k hands: gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools and Applications*, 2019.

[18] Adrian Bulat and Georgios Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *International Conference on Computer Vision*, 2017.