MyStyle: A Personalized Generative Prior

Yotam Nitzan[1,2] Kfir Aberman[1] Qiurui He[1] Orly Liba[1] Michal Yarom[1]
Yossi Gandelsman[1] Inbar Mosseri[1] Yael Pritch[1] Daniel Cohen-Or[2]

[1] Google Research [2] Tel-Aviv University

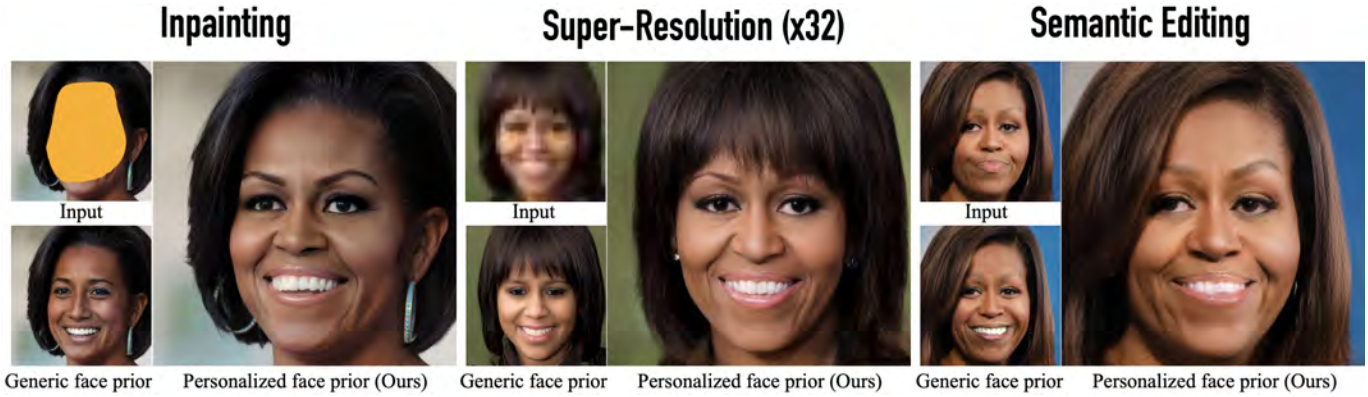


Figure 1. Using our personalized prior tuned with images of Michelle Obama, we solve various challenging tasks while faithfully preserving her key facial characteristic. Left to right: inpainting, super-resolution, and semantic editing (smile). Each example shows the original input image of Obama, which may be corrupted (top left), and the output based on our personalized face prior (right), compared to a generic face prior (bottom left). The generic face prior is learned from a diverse set of images and produces results that do not preserve Obama's key facial characteristics.

Abstract

We introduce MyStyle, a personalized deep generative prior trained with a few shots of an individual. MyStyle allows to reconstruct, enhance and edit images of a specific person, such that the output is faithful to the person's key facial characteristics. Given a small reference set of portrait images of a person (~ 100), we tune the weights of a pretrained StyleGAN face generator to form a local, low-dimensional, personalized manifold in the latent space. We show that this manifold constitutes a personalized region that spans latent codes associated with diverse portrait images of the individual. Moreover, we demonstrate that we obtain a personalized generative prior, and propose a unified approach to apply it to various ill-posed image enhancement problems, such as inpainting and super-resolution, as well as semantic editing. Using the personalized generative prior we obtain outputs that exhibit high-fidelity to the input images and are also faithful to the key facial characteristics of the individual in the reference set. We demonstrate our method with fair-use images of numerous widely recognizable individuals for whom we have the prior knowledge for a qualitative evaluation of the expected outcome. We evaluate our approach against few-shots baselines and show that our personalized prior, quantitatively and qualitatively, outperforms state-of-the-art alternatives.

Additional results and information are available on our website.

1. Introduction

Our personal digital album contains a myriad of images depicting ourselves in different scenery, poses, expressions, and lighting conditions. Although each image can be very different from the others, they all contain our unique facial characteristics. Leveraging this property of one's photo collection may enable the development of editing operations that are tailored to a specific individual, providing faithful reconstruction of their key facial characteristics in various scenarios. In particular, such an approach may be useful in image enhancement applications where only partial cues related to perceived identity of the captured subject are present, such as super-resolution, deblurring, inpainting, and more.

In recent years, the domain of image editing and enhancement in general, and face editing in particular, has experienced a significant shift. From pixel-level editing approaches [6], the field gradually shifted to latent-space editing methods that essentially interpret operators that are applied in a latent space of a generative model [19] as explicit image editing operations. This new latent-space-based editing enables new capabilities and demonstrates state-of-the-art performance. In particular, StyleGAN [27] became the gold-standard and core component for intuitive editing of face images [2, 53]. While results are impressive, all methods operate by hallucinating information from a general *domain prior* that is learned from a large and diverse dataset containing many identities. Hence, when editing

an image of a recognizable person, such as Michelle Obama, the result of using a generic face prior may be a person that only resembles Obama and does not preserve her key facial features (see Figure 1).

In this work, we propose an approach to a new problem - *personalization* of a face prior, which addresses the following question: Given a few portrait images of an individual, can we learn a *personalized-prior* that facilitates face-editing and enhancement operations while being faithful to the unique facial characteristics of that individual?

A naive approach to obtain a personalized generative prior, would be to apply a few-shot training method [34, 60], on a given set of images of a person. However, as we demonstrate in the paper, these techniques do not adequately preserve the key identity of an individual. We speculate that one fundamental reason is that these methods map a distribution of a few images into the entire latent space, and that this "smearing" effect degrades the quality and preservation of generated fine key facial characteristics.

In contrast, our proposed few-shot tuning technique, illustrated in Figure 2(a), affects only a compact low-dimensional, local, manifold in the latent space. Our tuning extends PTI [41] from capturing one image with a single latent, to capturing an individual's identity using a manifold. The manifold embeds the individual's images and constitutes as a personalized prior that enables generation of identity features that are faithful to the individual depicted in the images. In addition, the use of a personalized prior enables mitigation of possible biases in the diversity of the training data towards specific attributes (e.g., skin tone) [13], that may lead to undesired and non-inclusive behavior of enhancement approaches [32].

In practice, we begin with a generator pretrained on a dataset of general, diverse, faces [27]. Such a generator constitutes a *domain prior* which encapsulates understanding of high-quality face imagery, semantic facial features, and more. We aim to preserve these properties while tailoring the prior for a specific person. Given a small set of images of a particular individual, we next tune the generator's weights, such that each image in this set is reconstructed with a particular fixed code, coined *anchor*. The anchor is calculated by projecting the image onto the latent space using a dedicated encoder. Because our tuning is applied to specific regions in the latent space, it only affects a low-rank local manifold, enabling preservation of the essential properties of StyleGAN.

Owing to the latent space being smooth and disentangled, linear combinations of anchors are also modified to be faithful to the identity of the individual and exhibit fusion of low-level and high-level attributes that appear in the reference set. In particular, we show that the convex hull defined by the anchors constitutes a *personalized prior*.

Finally, we leverage this newly created personalized prior and propose a dedicated projection method for various image enhancement tasks, as well as a novel method for identity preserving semantic editing.

Our contribution is threefold: (i) we introduce the notion of a personalized face prior; (ii) we present a few-shot tuning technique to obtain a personalized subspace in the StyleGAN latent space, and (iii) we propose a unified approach that leverages it for personalized image enhancement and editing applications and achieves state-of-the-art identity preserving results.

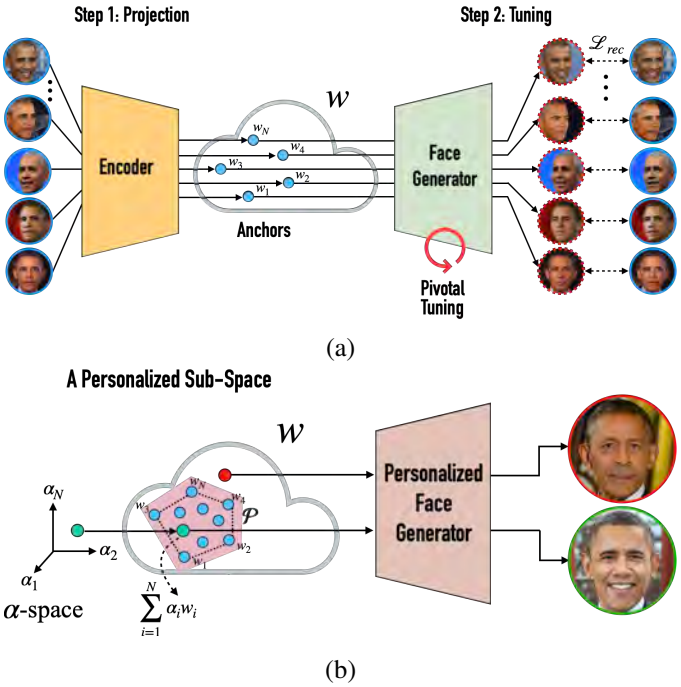


Figure 2. Creating a personalized, local and low-dimensional sub-space in the latent space of a face generator. (a) Given a set of N portrait images of an individual, we first project them into StyleGAN's $\mathcal{W}$ space with a pretrained encoder to extract a set of fixed codes we refer to as *anchors*, $\{w_i\}_{i=1}^{N}$. The anchors correspond to the set's nearest possible neighbors in latent space. We then tune the weights of a pretrained StyleGAN generator such that each anchor reconstructs its corresponding image from the set. (b) The set of the anchors' coefficients $\{\boldsymbol{\alpha}_i\}_{i=1}^{N}$, defines a new space, coined α-space– a low-dimensional space that is mapped into a personalized sub-space $\mathcal{P} \subset \mathcal{W}$. $\mathcal{P}$ contains latent codes associated with diverse head-shots of the individual, unlike samples outside of $\mathcal{P}$ that may generate different identities.

We extensively evaluate our framework for several different applications and against multiple baselines. In particular, we evaluate our method as a few-shot generator that can synthesize realistic faces that are faithful to the specified identity, as well as a personalized prior that regularizes image inpainting, super-resolution and semantic editing tasks. Additionally, we perform a thorough analysis of the personalized space, enabling detailed understanding of its key qualities. With qualitative and quantitative evaluation, and a user study, we demonstrate that our approach excels in capturing the characteristic features of the face identity and outperforms alternatives in all the aforementioned applications.

2. Related Work

2.1. Generative Prior

Learning image priors has been a long-standing goal of the computer vision and graphics communities as it is essential to solve a wide range of image inversion tasks [18, 22, 42, 48, 57].

Recently, pre-trained GANs [19] have become the image prior of choice for many popular tasks. Specifically, methods for image enhancement tasks such as inapinting, deblurring, and super-resolution [12, 20, 31, 32, 35, 39, 52, 56], commonly rely on a generative prior. These methods produce an enhanced image by projecting a corrupted input into the latent space of a GAN. Another task commonly approached using a generative prior is semantic editing. First, a real image is inverted into the latent space of the GAN [1, 3, 47]. Next, the resulting latent code can be semantically edited using a wide range of methods [2, 21, 36, 43, 49]. The edited latent code is then forwarded through the generator to produce the edited image. For a thorough introduction, we refer the reader to a recent survey [8].

A substantial issue with aforementioned works is that the projection at their core is often not accurate enough. The generative prior is inevitably unable to faithfully represent all images of the domain. This problem is often mitigated by using an enlarged latent space [1]. However, this is a mixed blessing, as this approach is known to weaken the prior [47], presenting a trade-off. Recently, several works have proposed methods [4, 7, 35, 41] to overcome this trade-off by changing the prior slightly so it would include a representation of a given image. After the generator has been trained on the large and diverse set, these methods tune the generator further to reconstruct a single image on which they operate. To prevent over-fitting, the training procedure is regularized by limiting it to a small set of weights [4, 7], progressively opening weights for training [35] and preventing the generator from changing other outputs [41].

While the issue of expressivity received considerable attention, another fundamental issue has not yet been addressed. The generative prior described is learned from a large and diverse dataset, including many objects from the data domain. As such, it is a generative prior of the entire *domain*, but it lacks any notion of personalization. For example, as shown in Figure 1, inpainting an occluded face of a person would produce a plausible output, but it would not be the person who was originally captured in the image. In this paper, we propose a new task of creating a *personalized* generative prior, so such applications can be faithfully performed. To this end, we fine-tune a generator consisting of a domain prior to form this personalized prior. Differently from previous works [3, 7, 35, 41], we do not limit our method to tune the prior to capture a single image, but rather a rich distribution of appearances of an individual.

2.2. Few-Shot Generative Models

Forming a personalized generative prior first requires we train a generative model. Many previous works have proposed methods to train generative models with limited data. Such works can broadly divided to two – methods that require pre-training and methods that can be trained-from scratch, but also benefit from pre-training. A shared goal for all method is preventing over-fitting.

Methods that can train from scratch typically mitigate over-fitting by burdening the discriminator during adversarial training, for example by providing it with augmented data [26, 60] or an auxiliary task [55]. Such methods still require thousands of images but are fairly robust for training on a wide range of data domains. The transfer-learning works are able to train with extremely few images, as few as none [16]. They typically do so by regularizing training, either with a dedicated loss [30, 34] or by restricting which weights are trained [16, 33, 40]. We note that such works obtain great results on "artistic" domains, such as caricatures. However, when realism is key, as is the case for human faces, these methods are unable to produce satisfactory results.

We experiment with several few-shot methods and find that they all fail to adequately preserve the identity of an individual depicted in a few images. We speculate that one fundamental reason is that these methods map a distribution of a small set of images into the entire latent space, only enabling preservation of coarse features from the target domain. We therefore propose a different training method based on Pivotal Tuning [41], which proves to be superior for our setting.

2.3. Personalization

We are all fundamentally different. This fact has led to the production and development of items, tools and methods that are tailored for a specific individual. Examples exist in a wide range of domains from everyday life such as clothing, medicine and nutrition. In recent years, personalization has also become an important factor in some fields of Machine Learning research such as recommendation systems [5], language models [11] and Federated Learning [24]. Compared to these fields, personalization has not yet made as strong of an impact on Computer Vision and Graphics.

Facial image enhancement applications would benefit greatly from personalizing for a specific individual. By doing so, such methods will be able to produce more faithful and realistic results. Recently, several facial enhancement works [14, 17, 29, 51, 61] have provided a reference image of the subject at inference to serve as a visual cue. However, these methods are limited to relatively simple tasks and settings and their performance does not match those of their counterparts ignoring personalization.

In this work, we overcome this gap and bring personalization to the forefront of image enhancement and editing. We propose a new problem of forming a personalized generative prior and introduce a novel method to do so. Our method produces highly identity preserving results for a wide range of applications with comparable image quality to state-of-the-art non-personalized methods.

3. Method

Our objective is to create a *personalized generative prior* for an individual given relatively few photos of them, roughly 100.

To this end, we seek to adapt a StyleGAN [28] model, trained on a large scale face dataset, such as FFHQ [27], to capture a personalized prior on top of the domain prior.

Our approach can be divided into two main parts. First, we propose an intra-domain adaption method based on Pivotal Tuning [41] (Subsection 3.1). The tuning is applied to specific regions in the latent space, and hence, does not transform it uniformly. Once applied, we identify a low-rank, local, manifold that has been personalized while preserving the high quality typical to StyleGAN (Subsection 3.2). Second, we leverage this newly created personalized prior and propose a generic method for various image enhancement tasks based on latent space projection (Subsection 3.3), as well as a novel method for identity preserving semantic editing (Subsection 3.4).

3.1. Adaptation

Given a reference set of N images depicting an individual, $\mathcal{D}_p = \{x_i\}_{i=1}^N$, we aim to adapt a pre-trained StyleGAN generator, G_d, constituting a domain prior into one constituting a personalized prior, G_p. We denote $\mathcal{W}_d$, $\mathcal{W}_p \subseteq \mathbb{R}^k$, to be the learned latent spaces of G_d and G_p, respectively. Where k is the dimension of the latent space, 512 for StyleGAN.

To obtain G_p, we propose a training scheme inspired by Pivotal Tuning [41], shown in Figure 2(a). We aim to change $\mathcal{W}_d$ as little as possible so it would encode the personalized reference set $\mathcal{D}_p$, without harming the rich semantic domain prior previously learned. This change should incorporate the individual's identity without affecting other semantic attributes. To facilitate a minimal change, we first seek for $w_i \in \mathcal{W}_d$ that when passed through G_d reconstructs x_i as accurately as possible. Due to the limited expressiveness of StyleGAN in $\mathcal{W}$ space [1, 47], $G_d(w_i)$ is still different from x_i, and specifically in terms of the person's identity. We therefore tune G_d with a simple reconstruction objective – reconstruct x_i given the latent code w_i.

In practice, we take $\{w_i\}_{i=1}^N$, which we call *anchors*, to be the latent space inversions produced by a pretrained encoder [39]. For the tuning of the generator, we follow common practice and use a combination of a pixel loss, L_2, and an LPIPS loss [58], $\mathcal{L}_{lpips}$. Formally, the reconstruction loss for a single sample is

$$\mathcal{L}_{rec}(G, x_i, w_i) = \mathcal{L}_{lpips}(G(w_i), x_i) + \lambda_{L_2} \|G(w_i) - x_i\|_2. \tag{1}$$

where λ_{L_2} is a hyperpatameter balancing the two losses.

Finally, we aim to find G_p that minimizes the reconstruction loss over the dataset, namely,

$$G_p = \arg\min_G \mathbb{E}_i \left[\mathcal{L}_{rec}(G, x_i, w_i)\right], \tag{2}$$

where the generator is initialized with G_d.

We note that the tuning process optimizes a simple reconstruction loss with no regularization. Nevertheless, despite being trained on a small set, we empirically observe no over-fitting. We speculate that the optimization goal being local, highly contributes to this result. Common techniques to train a generative model, such as adversarial training, requires the generator to follow the data distribution from all input latent codes. The generator is thus forced to change on infinitely many, highly varying inputs. A simple solution to such a hard task is to reproduce the training set. Our tuning method on the other hand presents an easier task, the generator's weights should be changed only for a small number of inputs.

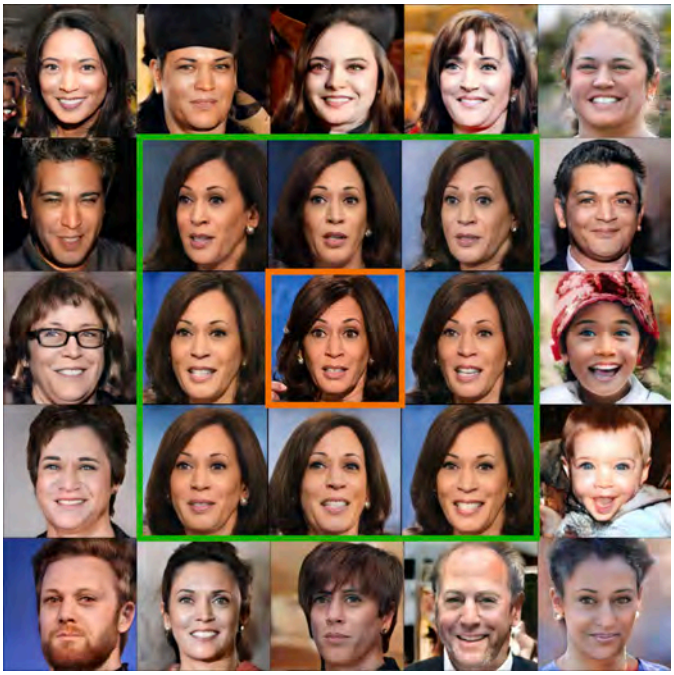


Figure 3. A personalized neighborhood. Samples of a generator tuned on 110 images of Kamala Harris. (i) An anchor (in the center, orange border), (ii) images synthesized by codes residing near the anchor (first ring, green border), (iii) Random synthesis in $\mathcal{Z}$ space (outer ring). It can be seen that images sampled around the anchor resemble the appearance and key facial characteristics of the person depicted in the anchor while other regions in the latent space depict a diverse set of faces.

3.2. Obtaining a Personalized Sub-Space

We have now obtained a generator G_p that reconstructs the reference set from the anchors. Naturally, a question arises – *what is encoded in other latent codes?* As we next explain, the answer fundamentally depends on the location of the latent code with respect to the anchors.

We empirically find that latent codes "close" to an anchor have also been personalized, *i.e.*, possess the identity of the individual (See Figure 3). This effect is expected since a GAN's generator is often smooth with respect to its input. StyleGAN's generator, in particular, was explicitly trained for this purpose [28].

While our training is completely local, optimizing only on a finite set of anchors, it empirically appears easier for the generator to maintain smoothness and propagate the effect to neighboring latents than doing otherwise. This "ripple effect" was considered harmful by Roich et al. [41], and they regularized training to prevent it. On the contrary, we willingly embrace it. Such neigh-

boring latents are **implicitly** trained to portray the individual's identity but not to reconstruct any image. Therefore, these latents expand the personalized prior and increase its expressive power.

On the other hand, latents that are "far" from all anchors are disrupted. Such latents are not constrained neither by an explicit loss nor by smoothness constrains, and are therefore victims of unexpected behavior. In practice, as can be seen in Figure 3, such latents often depict a minor level of personalization but at the cost of low image quality and realism.

Let us now consider that the latent space is not only smooth but also highly disentangled. Therefore, the linear interpolation between two latent codes rarely depict features that are absent from both endpoints [27]. While identity is a high-level, ambiguous feature, we find that this property still holds. Hence, identity is preserved during a linear interpolation between two anchors, as they depict the same identity. A simple induction then suggests that every convex combination of anchors is identity preserving.

Given both aforementioned findings - we consider a *"Dilated" Convex Hull* defined by the anchors to be a personalized subspace within $\mathcal{W}_p$. Intuitively, by dilated we refer to expanding the convex hull outwards to capture more neighboring latent codes.

Throughout this work, we find it convenient to represent the convex hull using normalized generalized Barycentric coordinates [15]. Simply put, let us consider $V = span(\{w_i\}_{i=1}^N)$. Then, generalized Barycentric coordinates are the coordinates with respect to the set of anchors as basis, whose sum is 1.

Using normalized generalized Barycentric coordinated, the β-dilated convex hull is easily defined by

$$\mathcal{A}_\beta = \{\boldsymbol{\alpha} \in \mathbb{R}^N | \sum_i \alpha_i = 1, \forall i : \alpha_i \geq -\beta\}, \tag{3}$$

where $\beta \in [0, \infty)$ controls the amount of dilation.

To translate the convex hull into the standard coordinates used for $\mathcal{W}_p$, a linear transformation is applied. Let $M \in \mathbb{R}^{k \times N}$ be the matrix with the anchors along its columns. Then the dilated convex hull is given by

$$\begin{aligned} \mathcal{P}_\beta &= \{M\boldsymbol{\alpha} | \, \boldsymbol{\alpha} \in \mathcal{A}_\beta\} \\ &= \{\sum_i \alpha_i w_i | \sum_i \alpha_i = 1, \forall i : \alpha_i \geq -\beta\}. \end{aligned} \tag{4}$$

From Eqn. (4) we note that, $\mathcal{P}_0$ is the set of all convex combinations of the anchors and is therefore by definition the convex hull of the anchors. Intuitively, one can consider the α-space to be the space of the coefficients of those convex combinations. On the other hand, $\mathcal{P}_\infty = V$. We often omit β from the notation and informally note $\mathcal{P}_\beta$ as $\mathcal{P}$ and $\mathcal{A}_\beta$ as α-space.

Dilating the convex hull captures more "close" points and expands the personalized space. However, dilating it too much leads to the inclusion of "far" points that were corrupted. Clearly, the distinction between "close" and "far" latents is a simplification. The actual effect of tuning is not discrete but continuous. Therefore, as β increases, $\mathcal{P}_\beta$ contains additional latent codes that are more expressive and diverse but are less faithful to the personalized prior. Meaning, β controls a tradeoff between the prior – image quality and personalization – and expressiveness. The value of β could be determined empirically, according to the user's preference between the prior and expressiveness.

We last note that $\mathcal{P}_\beta$ differs significantly from common latent spaces considered in literature, *e.g.* $\mathcal{Z}, \mathcal{W}, \mathcal{S}$ [53]. First, it is a manifold rather than a probability distribution or simply a Euclidean space. Second, as often the number of images, N, is smaller than the dimension of the original latent space, k – $\mathcal{P}_\beta$ is of low-rank. Last, and arguably most important, it is bounded. These differences affect several components of our method and we are often able to leverage them to obtain improved performance.

3.3. Personalized Image Enhancement

Having obtained and modeled a personalized prior. We next propose a novel projection method that can be used to leverage the prior for a variety of image enhancement tasks, such as inpainting and super-resolution as well as GAN inversion [8].

Several recent works [31, 32] have taken similar approaches for leveraging domain generative prior for image enhancement – projection into latent space. These methods are given a degraded image I_d and access to a differentiable function simulating the degradation, ϕ. They then devise methods to find the latent code from which StyleGAN produces an image, that after degradation with ϕ reconstructs I_d, most accurately. Formally, they seek for

$$w^* = \arg\min_w \mathcal{L}((\phi \circ G)(w), I_d), \tag{5}$$

where $\mathcal{L}$ is some reconstruction loss. To ensure the high fidelity of the projection, such works use the $\mathcal{W}^+$ space [1] which is virtually infinitely expressive. But for that reason, they also need to regularize the w^+ code, preventing the simple solution of degradation appearing in $G(w)$. Finally, they take the enhanced output image to be $I_e = G(w^*)$.

We next devise a method for projecting an image into our personalized prior, that adopts this common approach and adapts its different components. An illustration of the method is displayed in Figure 4.

In order to leverage the prior, one must restrict the optimization solution to remain on the manifold. This restriction can be efficiently implemented in α-space, and we therefore project to α-space, rather than to $\mathcal{P} \subset \mathcal{W}$. We now adapt the projection problem described in Eqn. (5) to use α-space. Adopting the reconstruction loss defined in Eqn. (1), the new projection problem can be described as

$$\boldsymbol{\alpha}^* = \arg\min_{\boldsymbol{\alpha}_\beta \in \mathcal{A}_\beta} \mathcal{L}_{rec}(\phi \circ G, I_d, M\boldsymbol{\alpha}). \tag{6}$$

The restriction for $\boldsymbol{\alpha}_\beta$ to remain in $\mathcal{A}_\beta$ is implemented intuitively and efficiently using the constraints in α-space's definition. To bound the minimal negative value of $\boldsymbol{\alpha}_\beta$ to $-\beta$, we pass an unrestricted $\boldsymbol{\alpha}$ through a softplus function shifted by β. Formally,

$$\boldsymbol{\alpha}_\beta = \frac{1}{s} log(1 + e^{s(\boldsymbol{\alpha}+\beta)}) - \beta. \tag{7}$$

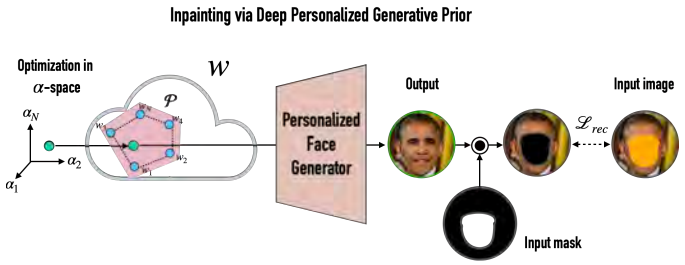


Figure 4. Personalized inpainting. Having obtained a personalized sub-space $\mathcal{P}$, we can use it as a generative prior to reconstruct occluded parts of a person in a given image such that the output preserves the key facial characteristics of the person. Given an input image and a mask, our framework optimizes a code in α-space to reconstruct the non-occluded parts of the input image.

Note that $\boldsymbol{\alpha}$ is a vector and all operations in Eqn. (7) are element-wise. s is a "sharpness" hyper-parameter.

To restrict the sum of $\boldsymbol{\alpha}_\beta$ to 1, we add both a soft penalty term to the optimization objective, given by

$$\mathcal{L}_{sum}(\boldsymbol{\alpha}_\beta) = \left(\sum_i \alpha_i - 1\right)^2, \tag{8}$$

and explicitly normalize the optimization result, as often done in such cases [37].

We have thus far described a projection method to α-space. Next, inspired by $\mathcal{W}^+$'s extended expressiveness, we similarly define a new latent spaces $\mathcal{A}^+_\beta$, coined α^+-space, that may use a different $\boldsymbol{\alpha}_\beta$ for each layer of the generator. The latent space $\mathcal{P}^+_\beta$, is defined similarly as before (see Eqn. (4)). The projection described for α-space is generalized to α^+-space, where $\mathcal{L}_{sum}$ is calculated per-layer and then averaged.

To regularize the solution in α^+-space, we adopt the regularization proposed in e4e [47] – to minimize variation between the latents in different layers. Technically, instead of optimizing a different latent code in each layer, we hold a single latent $\boldsymbol{\alpha}$ and an additive offset for each layer, *i.e.* $\boldsymbol{\alpha}^+_\beta = (\boldsymbol{\alpha}_\beta + \boldsymbol{\Delta}_0, ..., \boldsymbol{\alpha}_\beta + \boldsymbol{\Delta}_N)$. We then regularize the norm of the offsets $\Delta = \{\boldsymbol{\Delta_i}\}_{i=1}^N$, via

$$\mathcal{L}_{d-reg}(\Delta) = \sum_i \|\boldsymbol{\Delta_i}\|_2. \tag{9}$$

Our final objective is thus given by

$$\mathcal{L}_{final}(\alpha, \Delta, G, \phi, I_d, M) = \mathcal{L}_{rec}(\phi \circ G, I_d, M\boldsymbol{\alpha}^+_\beta) + \lambda_{d-reg}\mathcal{L}_{d-reg}(\Delta) + \mathcal{L}_{sum}(\boldsymbol{\alpha}^+_\beta). \tag{10}$$

We optimize the objective for $\boldsymbol{\alpha}$ and Δ

$$\boldsymbol{\alpha}^*, \Delta^* = \arg\min_{\alpha,\Delta} \mathcal{L}_{final}, \tag{11}$$

and the final enhanced image is taken to be $G(M\boldsymbol{\alpha}^{*^+}_\beta)$.

3.4. Personalized Semantic Editing

A popular application of generative priors is performing semantic editing in latent space. Commonly, this is performed by gradually moving an initial latent code, w, along a linear latent direction, $\mathbf{n}$. The edited latent code is then given by $w_{edit} = w + \theta\mathbf{n}$, where θ is a scalar that determines the magnitude of the edit step. The latent direction $\mathbf{n}$ controls a factor of variation that exists within the training set, in a disentangled fashion. We seek to perform personalized semantic editing. *I.e.*, semantic editing that is both identity preserving as well as typical to that individual.

Prior works [16, 38, 41, 54] have leveraged the natural alignment of fine-tuned generators [54] to apply editing directions located in a parent generator in the child's latent space. We empirically find that this property holds for our generator G_p as well. Therefore, the multitude of directions learned for generic-trained StyleGAN can be used with G_p as well.

However, these directions are clearly learned from the entire domain and are not personalized. Considering $\mathcal{P}$ is the personalized space, it is clear why the direction is not personalized. First, $\mathbf{n}$ could be any vector in $\mathcal{W}_p$, and specifically not limited to reside within V. Any small step in such a direction, would exit V and therefore $\mathcal{P}$. Second, assuming that by chance $\mathbf{n} \in V$, infinitely traversing it would inevitably stray away from $\mathcal{P}$, which would lead to degredations to quality and identity. The second issue, in fact, is not unique to our setting and generally exists in GANs and their latent spaces. Traversing "too far" along such directions inevitably leads to regions in which the probability density is low and causes degradation [45].

We next propose a method to prevent any linear direction, regardless of how it was obtained, from straying far from the personalized prior. The method is composed of two simple steps, each solving one of the issues previously discussed, and thus ensuring the edited latent remains in $\mathcal{P}$. We first aim to express $\mathbf{n}$ in α-space coordinates. As $\mathbf{n}$ is not restricted to V, this might be impossible. We instead express the α-space coordinates of its projection $proj_V(\mathbf{n})$, given by $\gamma = (M^TM)^{-1}M^T\mathbf{n}$. The initial latent code, w, is obtained using the projection method described in Subsection 3.3, to invert a real image. Therefore, it is already given in α-space coordinates – α_w. Now, we can easily transform the linear editing in $\mathcal{W}_p$ to a linear editing in α-space,

$$w_{edit} = w + \theta\mathbf{n} = M(\alpha_w + \theta\gamma). \tag{12}$$

This completes the first step, the editing is now restricted to V.

Next, we evaluate how far does w_{edit} stray away from $\mathcal{P}_\beta$ by measuring the minimal β-dilation that includes it. This is trivially done by computing $\beta_\theta = |\min(\alpha_w + \theta\gamma)|$. As β_θ increases, the editing strays further from $\mathcal{P}$ and gradually the prior weakens. We now give the user a choice. Assuming a maximal value of desired β-dilation is known, one can stop editing once it is reached. Essentially, this operation transforms an infinite linear editing to one with endpoints. Endpoints for semantic editing were previously promoted by Spingarn-Eliezer et al. [45] to prevent straying from the dense regions in latent space.

Alternatively, after the maximal β-dilation is reached, one can continue editing and mitigate the deviation from the prior by slightly trading-off disentanglement. Assume one had reached a w_{edit} outside of the desired dilation, $\mathcal{P}_\beta$. One can project w_{edit} back to $\mathcal{P}_\beta$, by solving a simple convex optimization problem. While this maintains w_{edit} with the desired prior, it also causes it to deviate from the linear direction $proj_V(\mathbf{n})$, and therefore might affect other properties and degrade disentanglement. In practice, we find the projection useful as it allows to extend the range of editing while preserving identity and quality. Additionally, since $\mathbf{n}$ is not perfectly disentangled, we often find that the degradation to disentanglement caused by projection is acceptable.

4. Experiments

We next turn to investigate several components and properties of our method and data and their effect on the personalized prior. Additional experiments appear in Section C.

4.1. Evaluating the Locality of the Prior

One of the unique and useful qualities of $\mathcal{P}$ is the fact that it is local. This property is visualized in Figure 3, where it is demonstrated that latent codes within $\mathcal{P}$ are personalized, while latent codes sampled randomly from $\mathcal{W}_p$ are not.

We next perform a simple experiment to quantitatively support these findings. We perform a linear interpolation from $\mathcal{P}$'s center, c, to a random anchor, w_i, and then further extrapolate in that direction. Note that the interpolation is entirely within $\mathcal{P}_0$ while the extrapolation is straying away from it, corresponding to increasing β values. We aim to quantify how identity preservation is effected along this traversal.

To this end, we follow common practice [58] and leverage similarity of deep features to measure faithfulness to the personalized set. Specifically, we measure the cosine similarity between the generated image's features and the nearest-neighbor features from the reference set. We refer to this metric as Maximal Perceptual Similarity (MAPS). The deep features may be extracted from various deep neural networks. In our experiments, we leverage a classifier [44] trained on the domain.

In Figure 5 we present results for several such traversals, as well as one traversal between the same latents but in the FFHQ generator. As expected, identity is strongly preserved for latents in $\mathcal{P}_0$ and becomes stronger in the anchor's close proximity. After reaching the anchor, the trend changes. As the traversal strays further away from $\mathcal{P}_0$, corresponding to increasing β values, identity preservation gradually vanishes.

4.2. Latent Spaces of G_p

We next evaluate the effect the choice of latent space has on image enhancement results. As demonstrated in Figure 6, projecting to $\mathcal{P}_\beta^+$ is favorable for image enhancement. Enhancement results in $\mathcal{W}_p$ and $\mathcal{W}_p^+$ are of low quality and not personalized. This support our previous findings that the space has not been personalized uniformly and additionally indicates that the projection does not converge to the personalized space without explicit regularization.

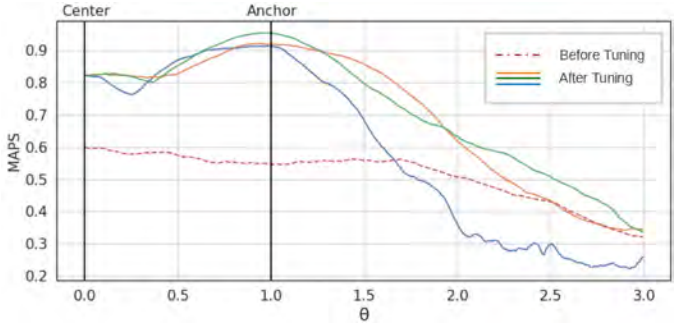


Figure 5. Quantitative evaluation of the locality of personalization. We perform a linear interpolation and extrapolation from $\mathcal{P}$'s center, c, to and beyond a random anchor, w_i. The traversal is given by $x = \theta w_i + (1-\theta)c$. We then report the MAPS score of images generated along this traversal as a function of θ. Each solid lines represent images generated from G_p with a different anchor. The dashed line represents images generated by G_d and is provided for reference. As can be seen, personalization is strong in $\mathcal{P}$ $(0 \leq \theta \leq 1)$ and degrades as the extrapolation strays further from $\mathcal{P}$. At roughly $\theta = 1.75$, the extrapolation reaches sparse regions of $\mathcal{W}$ and ceases representing realistic faces in all generators. This effect is especially noticeable in G_d, where until that point, the faithfulness of identity is relatively constant.

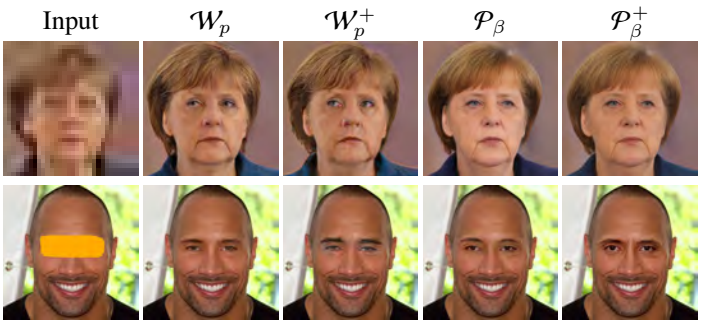


Figure 6. We demonstrate the effect of projecting degraded input images into different latent spaces for image enhancement. We find that projecting to $\mathcal{W}_p$ and $\mathcal{W}_p^+$ yields non-personalized and low quality results. Projection to $\mathcal{P}_\beta^+$ is superior to $\mathcal{P}_\beta$ in terms of personalization and fidelity.

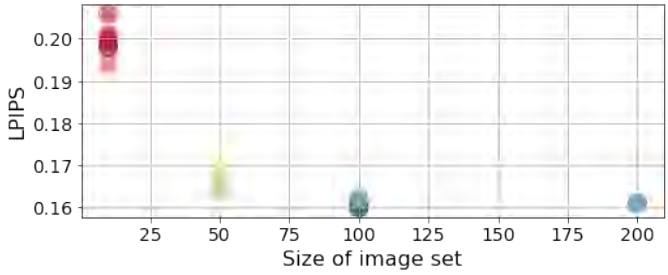
We additionally find that for image enhancement, $\mathcal{P}_\beta^+$ is superior to $\mathcal{P}_\beta$ in two manners. First, it provides more expressive power, as $\mathcal{W}^+$ does with respect to $\mathcal{W}$ (see first row in Figure 6). Second, we find that $\mathcal{P}_\beta$ sometimes produces slightly less identity preserving results than $\mathcal{P}_\beta^+$ (see last row in Figure 6). This is initially surprising as $\mathcal{W}$ is known to provide more reliable prior than $\mathcal{W}^+$, however with $\mathcal{P}_\beta$ and $\mathcal{P}_\beta^+$ the opposite appears to be true. We empirically find that results from $\mathcal{P}_\beta$ often fully leverage the allowed dilation and arrive at exactly β, while results from $\mathcal{P}_\beta^+$ often converge at a smaller dilation. We speculate that results in $\mathcal{P}_\beta$ stray to further dilation to mitigate the limited expressiveness.

4.3. Effect of Dataset Size and Diversity

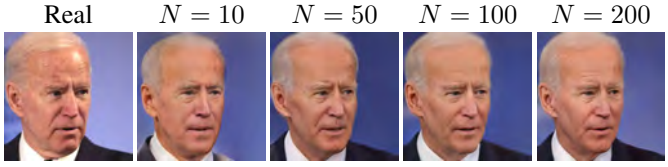
We now study the effect the reference set's size and diversity has on the quality of the personalized prior and generator. To this end, we measure the inversion accuracy of unseen test images into $\mathcal{P}_0$. Accurate inversions exists even in a non-personalized $\mathcal{W}^+$. However, since we restrain the inversion to $\mathcal{P}_0$, its accuracy provides an estimate to the expressive power of the personalized prior.

We sample several subsets of images from the reference sets of three individuals: Joe Biden, Emilia Clarke and Michelle Obama. The subsets are of sizes 10, 50, 100 and 200. For each subset, we estimate the diversity by computing the average pairwise LPIPS [34]. Additionally, we tune G_p and invert 20 test images following the projection protocol in Section 3.3. We then measure the reconstruction accuracy using LPIPS [58]. We repeat this experiment 5 times for all set sizes other than 200, which represents the size of the entire set in this experiment.

As can be seen for in Figure 7(a), at first, increasing the set size improves both the set's diversity and inversion accuracy. However, this is not the case for 100 and 200 images where an interesting phenomenon occurs. Although differences are relatively minor, performance correlates to the diversity of the set, regardless of the set size. While further experiments are required, we speculate that adding an image that does not contribute to diversity might add a burden to tuning and increase the dimension of α-space, thus hurting results. Visual sample of results is provided in Figure 7(b). One can observe the improvement from 10 to 50 and then 100, and then no major difference to 200.



(a)



(b)

Figure 7. The effect of reference set size and diversity on the prior's expressiveness. We sample subsets of different sizes from the reference set of Joe Biden. For each subset we additionally tune a model, G_p, and invert a set of test images to its $\mathcal{P}$. (a) Reports the inversion error using average LPIPS distance as a function of set size and diversity. Diversity is computed using average pair-wise LPIPS distances [34] and is reported as color in the spectrum between red (low) and purple (high). (b) Visual examples of inverting a given real image with various set sizes.

Table 1. Quantitative evaluation of few-shots synthesis approaches. Ours, Ojha et al. [34], and Diff-Augment [60]. The user study values are the percentages of images that appeared real to the users.

Method	User % (↑)	MAPS (↑)	Diversity (↑)
Ojha et al.	1.4	0.53 ± 0.08	2.42 ± 0.16
DiffAugment	31.7	0.76 ± 0.05	2.99 ± 0.17
MyStyle (Ours)	**68.9**	$\mathbf{0.79 \pm 0.04}$	$\mathbf{3.44 \pm 0.16}$
Real Images	83.1	1	0

5. Applications

In this section, we demonstrate the application of our personalized prior for popular generative tasks - image synthesis (Section 5.1), image enhancement (Section 5.2) and semantic editing (Section 5.3). In all experiments, we tune the generator from a pretrained FFHQ StyleGAN2 [28] on a personalized reference set. See Section A for more information regarding the datasets.

We note that our personalization approach could be considered as solving a few-shot domain adaption task from the domain of all faces to the domain of the face of a specific person. While existing domain adaptation works mostly focus on synthesis, their obtained generator can similarly be leveraged as a prior for all discussed applications. We therefore consider such methods as the most direct baseline and compare to them for all applications. For image enhancement we additionally compare to state-of-the-art methods designed for the specific application.

5.1. Image Synthesis

We compare our synthesis results with existing few-shot domain adaptation methods - Ojha et al. [34] and DiffAugment [60].

We initialize the generator to the same StyleGAN2 [28] model trained on FFHQ [27] and run each method's training approach on three personalized datasets: Adele (109 images), Kamala Harris (110 images), and Joe Biden (206 images). We then compare the synthesis of random images from these models.

To synthesize an image using our method, we sample a latent code from $\mathcal{P}_0$ and forward it through G_p. A simple protocol for sampling from $\mathcal{P}_0$ is described in Section B.1. For DiffAugment, we truncate the sampled latent code with $\psi = 0.7$ [10], which improves its results significantly.

Samples of generated images are displayed in Figure 8. Additional results are provided in the supplementary material.

We next evaluate the quality and diversity of the methods. The term *quality* refers not to the visual quality of the images, but whether their distribution is faithful to the training distribution. Therefore, in our setting high quality synthesis should, among other things, preserve the identity of the individual. We evaluate quality using two complimentary approaches.

First, we use MAPS, the perceptual similarity metric presented in Section 4.1. We note that similar to other quality metrics (*e.g.* FID [23]), MAPS is reported with respect to the

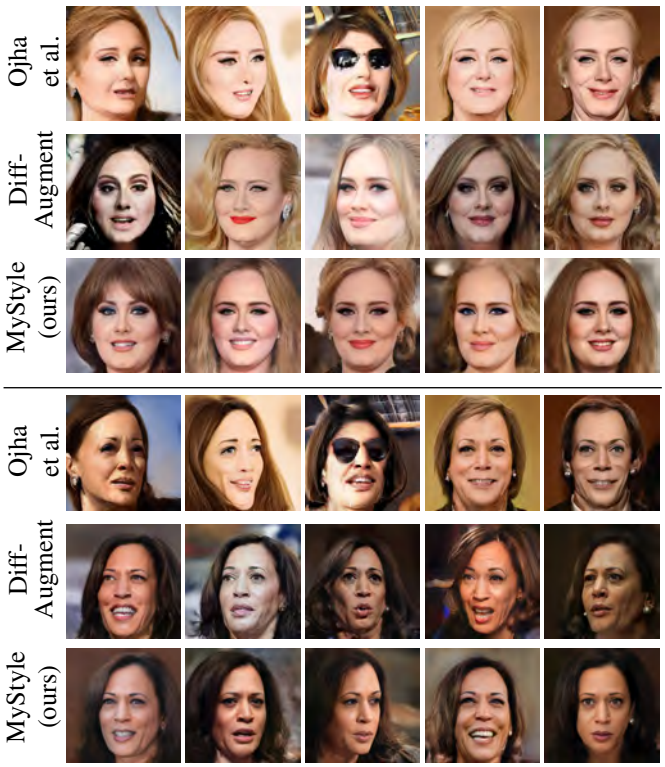


Figure 8. Random images synthesized by different approaches. In each block, (i) The few-shots method of Ojha et al. [34] (top row), (ii) Diff-Augment [60] (middle row), (iii) Our tuned generator sampled in the personalized sub-space (bottom row). As can be seen, our results are more realistic and identity preserving than those of the other methods.

training set. Second, we conduct a user study. Users were presented with an image which was either real or synthesized by one of the methods and were asked to choose whether it looks like a real image of the specific person. We asked users to respond to the survey only if they were familiar with what the person looked like. We gathered 1674 responses from 45 unique users.

For diversity, we follow the protocol suggested by Ojha et al. [34]. We synthesize 10,000 images from each model and cluster them according to their nearest neighbor in the training set. We then compute the mean and standard deviation of the intra-cluster LPIPS [58] distances.

Aforementioned metrics are averaged across all individuals, and reported in Table 1. Our model consistently outperforms both alternatives on all metrics. Qualitative samples are presented in Figure 8 and in our supplementary material. As can be seen, our results are significantly more realistic and identity-preserving. We specifically point the reader to observe the diversity in Adele's appearance which faithfully represents her different appearances over the years.

5.2. Image Enhancement

We choose the tasks of image inpainting and super-resolution as representative examples for image enhancement. We follow the same evaluation protocol for both. As competing methods, we use the generator obtained by DiffAugment [60] as an alternative prior, a state-of-the-art domain prior method and a version of it fine-tuned on a personalized reference set. The domain prior method serves only as a reference and represents existing methods. To compare all methods, we first present qualitative results for the reader's inspection, with additional results available in the supplementary materials. Next, we evaluate quality using MAPS and report user study results that reflect preference based on quality and fidelity.

In the study, users were presented with an input image and two results, one of ours and one of a baseline. They were then asked to pick the result which better resembles the person and has higher fidelity to the input. Results are reported as the percentage of responses that preferred a different method over ours. The images used in the user study are a randomly sampled subset of those used for quantitative evaluation.

We explicitly note that we do not evaluate quality based on reconstruction to a ground truth, neither quantitatively (*e.g.* L_2, PSNR and LPIPS) nor qualitatively. By definition, image enhancement methods hallucinate missing details, which might be valid despite differing from a specific ground truth image. Therefore, quantitative evaluation of reconstruction is meaningless and including ground truth in qualitative results might bias readers. As the individuals appearing in the experiments are most likely recognized by readers, we believe quality can be evaluated from the input-output pair alone.

5.2.1 Inpainting

The goal of image inpainting is to complete missing regions of an image. The degradation transform is modeled as multiplication of the original image with a binary mask m, *i.e.* $\phi(x) = x \odot m$ where $\odot$ is Hadamard's product. For the state-of-the-art domain prior method we use CoModGAN [59], which we also fine-tune and note by "CoModGAN + FT". Following common practice in literature, all methods assume m is known and let the result be a blend of the network's output in the missing region and the original image otherwise. We compare all methods for Barack Obama (192 images), Lady Gaga (133 images), and Jeff Bezos (114 images). We use $\beta = 0.02$ in all experiments.

Qualitative results are displayed in Figures 9 and 11. As can be seen, our method generates image completions that are more faithful to the person's identity. Importantly, even subtle features are restored, such as Jeff Bezos' drooping upper eyelid and Barack Obama's mole to the left of his nose. This is further supported by quantitative results reported in Table 2, where our results obtain higher MAPS scores and are strongly preferred by users.

5.2.2 Super-Resolution:

In super-resolution, we start from a low-resolution image $I \in \mathbb{R}^{3\times H\times W}$ and generate a corresponding high-resolution image $I \in \mathbb{R}^{3\times fH\times fW}$ where f is the upsampling factor. In this case, the degradation transform ϕ is downsampling the input image

Figure 9. Personalized inpainting qualitative evaluation. Left to right: input, CoModGAN [59], CoModGAN fine tuned on the reference set, DiffAugment [60], and our method. We suggest zooming-in to notice subtle differences in key facial characteristics.

Table 2. Quantitative evaluation of inpainting approaches: CoModGAN [59] trained on FFHQ, CoModGAN fine-tuned on our personalized reference set, our proposed inpainting approach with DiffAugment's generator as a prior [60], and MyStyle (Ours). The user study values reflect the percentages of responses (overall 430) in which the compared method was preferred over MyStyle.

Method	User % (↑)	MAPS (↑)
CoModGAN	0.9	0.55 ± 0.08
CoModGAN + FT	24.9	0.71 ± 0.08
DiffAugment	9.3	0.68 ± 0.09
MyStyle (Ours)	-	$\mathbf{0.72 \pm 0.08}$

by an $f \times f$ area kernel. We use GPEN [56] as the state-of-the-art baseline, which we also fine-tune and note by "GPEN + FT". All methods get a 32×32 input. DiffAugment and our method perform 32x upsampling while GPEN baselines perform an easier 16x upsampling due to the resolution of the official model uploaded by authors. We post-process the model output by replacing the non-face regions in the model output, segmented by Wadhwa et al. [50], with a Lanczos-upsampled version of the input image. We compare all methods for Michelle Obama (279 images), Emilia Clarke (258 images) and Xi Jinping (92 images). We use $\beta = 0.05$ in all experiments.

Qualitative results are displayed in Figures 10 and 11. As can be seen, our results are significantly more faithful to the person's identity, have comparable fidelity and superior visual quality. This is also demonstrated by quantitative results reported in Table 3, where our results obtain higher MAPS scores and are strongly preferred by users.

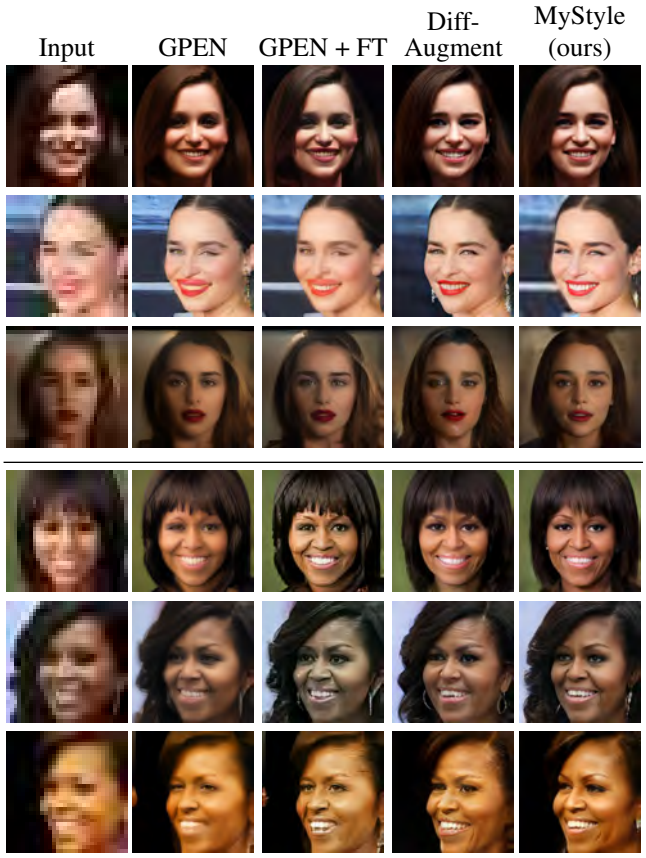


Figure 10. Personalized super-resolution qualitative evaluation. Left to right: input, GPEN [56], GPEN fine-tuned on the reference set, DiffAugment [60], and ours. Zoom-in to notice subtle differences in the key facial characteristics.

5.3. Semantic Editing

We now evaluate the performance of our personalized prior on semantic editing applications. Note that semantic editing is a multi-step process involving finding semantic linear directions, inverting a real image into latent space, applying the editing operator on the inverted latent code, and finally generating an image. In this section, we aim to compare only the process of applying the editing operator in latent space. To this end, we use the same linear directions and inversion method for MyStyle as well as the baselines. We use two InterFaceGAN [43] directions, for pose and smile, and PTI [41] as the inversion method.

For the state-of-the-art domain prior baseline, we apply semantic editing in a StyleGAN trained on FFHQ, as done by PTI.

Figure 11. Additional results of our personalized prior applied to inpainting, super-resolution and semantic editing, for widely recognizable individuals. Additional results are available in the supplementary material. We suggest zooming-in to better view fine details.

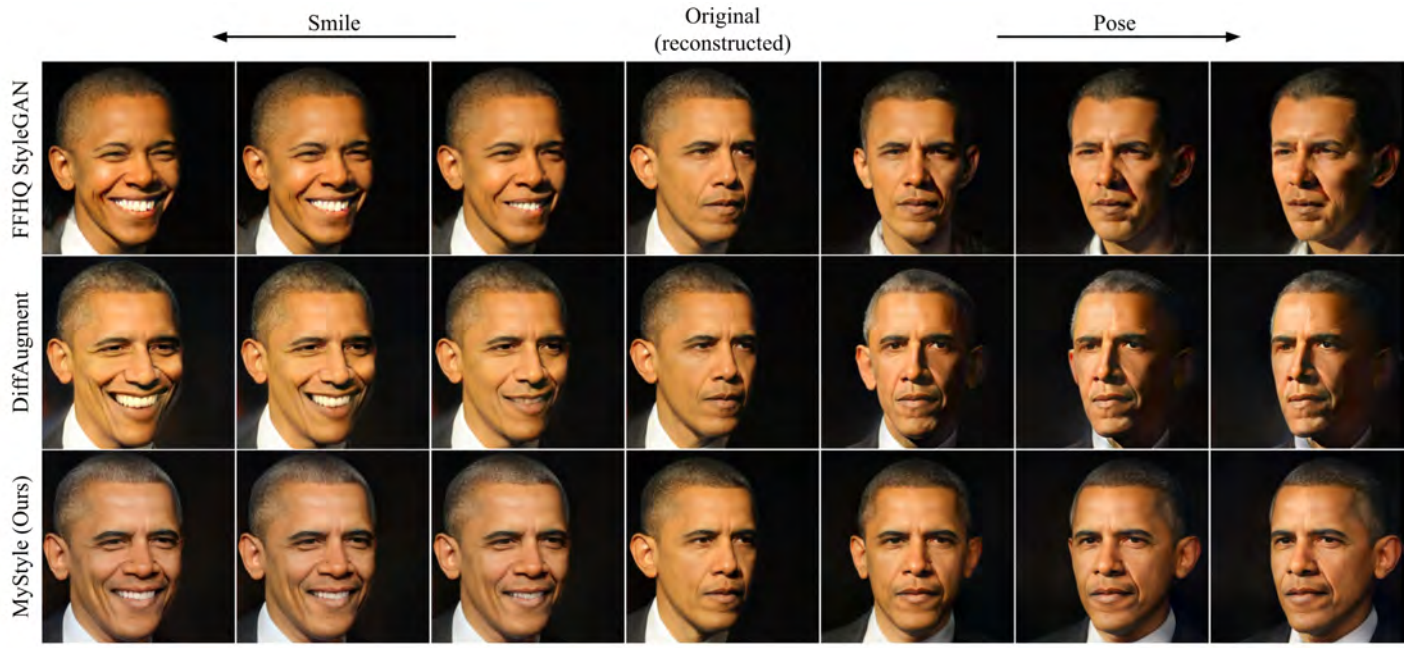


Figure 12. Personalized semantic editing. We use the same inversion method - PTI [41] and editing directions [43] with three different StyleGAN generators - FFHQ-trained (equivalent to PTI), fine-tuned with DiffAugment [60] and that obtained with MyStyle. The reconstructed input image is in the middle column, with smile editing to its left and pose editing to its right. Both competing approaches exhibit identity drift, caused by the entanglement of the edited attributes to identity features. In contrast, our approach preserves the key facial characteristics of Obama throughout the different editing outputs.

Table 3. Quantitative evaluation of super-resolution approaches: GPEN [56] trained on FFHQ, GPEN fine-tuned on our personalized reference set, our proposed super-resolution approach with DiffAugment's generator as a prior, and MyStyle (Ours). The user study values reflect the percentages of responses (overall 366) in which the compared method was preferred over MyStyle.

Method	User % (↑)	MAPS (↑)
GPEN	0.3	0.65 ± 0.09
GPEN + FT	2.2	0.75 ± 0.09
DiffAugment	4.6	0.75 ± 0.05
MyStyle (Ours)	-	$\mathbf{0.81 \pm 0.04}$

(a) Input (b) Domain edit (c) Personalized edit (d) Reference

Figure 13. Demonstration of the fundamental difference between editing with a domain prior and a personalized prior. The input image of Angela Merkel (a), is edited using a domain prior learnt from a large and diverse face dataset (FFHQ) (b). As can be seen, the edit maintains Merkel's coarse features. However, the smile is uncharacteristic of Merkel. This is because the notion of a smile was learned from thousands of different individuals. Conversely, when editing the expression using our personalized prior (c), the smile is more typical of Merkel, as demonstrated in the reference image (d).

Note that this method has no trained weights, other than those of the GAN, and therefore it cannot be fine-tuned. Comparison is made for Barack Obama (192 images) and Joe Biden (206 images).

A sample of qualitative results are portrayed in Figure 12. As can be seen, our method preserves identity for both smile and pose, whereas both alternative methods fail to preserve identity. See the supplementary material for more results.

We explicitly note that similar to previous applications, semantic editing using a domain prior is inherently unable to preserve identity. Nevertheless, it is common for methods that leverage a domain prior to aim at identity-preserving editing [2, 3, 41].

Consider adding a smile to an individual's face. These methods point to the lack of change in any unrelated factors of appearance (*e.g.* eyes or hair) as a sign of identity preservation. Even if a perfect semantic disentanglement is obtained, it is still not truly identity-preserving. The missing piece is that the smile itself should belong to the individual. In Figure 13 for example, there are two images of Angela Merkel for which a smile was added. Both edits are consistent with Merkel's general appearance. However, clearly, only one image portrays a realistic image of Merkel smiling. This is because, the smile added to Merkel's face is her own, as learned by our personalized prior.

We next quantitatively evaluate the performance of the meth-

ods for pose editing. For every input and method, we create a large gallery of edited images with head pose varying between $(-40, 40)$ degrees [62]. We then sample 5 equally spaced images. Quantitative results are reported in Table 4. As can be seen, our results are more identity-preserving and strongly preferred by users.

Table 4. Quantitative evaluation of editing with different priors. We compare editing of yaw angle in an FFHQ-StyleGAN domain's prior, DiffAugment generator's prior, and MyStyle's personalized prior (Ours). The user study values reflect the percentages of questions in which the compared method was preferred over MyStyle.

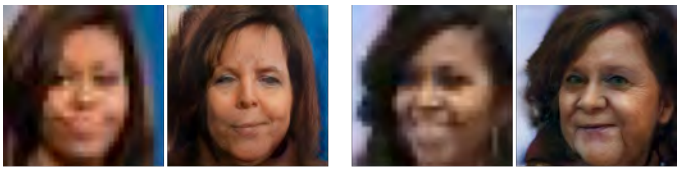
Method	User % (↑)	MAPS (↑)
FFHQ StyleGAN	2.3	0.60 ± 0.08
DiffAugment	10.9	0.66 ± 0.05
MyStyle (Ours)	-	$\mathbf{0.74 \pm 0.04}$

6. Conclusions

We have presented a new problem: forming a personalized prior, and introduced a method to achieve it through a few-shot tuning method. Our method takes a small sized (∼100-200) reference set of photos, and learns a personalized prior represented in a subspace of StyleGAN latent space. All style codes within that subspace represent images containing the individual's identity. We showed that this learned personalized prior promotes non-trivial performance for otherwise ill-posed tasks, and developed a mechanism to project an image onto the subspace using a novel latent space, α-space, that enable faithful reconstruction of the key facial characteristic of the person.

One noted limitation of the current approach is a direct result of the inherent limitation of StyleGAN that can not faithfully reconstruct images that are out of the training set's distribution while simultaneously maintaining them in "healthy" regions of the latent space, such images can include faces in extreme poses or faces that are occluded partially by accessories. Moreover, a mismatch between the identity of the person in the degraded image to the person depicted by the prior may break the realism of the output, resulting in a person that does not adequately resemble any of the individuals as demonstrated in Figure 14. It should be noted that automated, scalable image editing or photo manipulation methods must be researched and developed responsibly and consider fairness and content quality risks for potential downstream users.

An interesting direction for future work is to fine tune the generative prior to account for multiple individuals or for the aging factor and the temporal axis, where say, an individual has different hairstyles. We also believe that such restricted sub-spaces, can be effective for more tasks, beyond identity preservation, and can help direct or regularize various tasks, such as articulation of human poses, expressions of face, or shapes of chairs.



(a) (b)

Figure 14. Failure case - enhancement of degraded images of Michelle Obama (left) using a prior learned on Angela Merkel's reference set. It can be seen that due to the mismatch between the key facial characteristics of the person depicted in the reference set and the attributes of the person in the degraded image, the output images (right) contain visual artifacts and depict a person that does not adequately resemble any of the individuals.

Acknowledgments: We thank Rinon Gal, Or Patashnik, Amit Attia, Yuval Alaluf, Assaf Shocher, and Michael Rubinstein for reviewing early drafts and suggesting improvements. We thank Yogev Nitzan for his encouragement and Wei Xiong for help running ComodGAN baselines. This work was partially supported by the Israeli Science Foundation (3441/21, 2492/20).

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the StyleGAN latent space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 3, 4, 5, 17

[2] Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. StyleFlow: attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows. *arXiv preprint arXiv:2008.02401*, 2020. 1, 3, 12

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. ReStyle: A residual-based StyleGAN encoder via iterative refinement. *arXiv preprint arXiv:2104.02699*, 2021. 3, 12

[4] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. *arXiv preprint arXiv:2111.15666*, 2021. 3

[5] Justin Basilico Ashok Chandrashekar, Fernando Amat and Tony Jebara. Artwork personalization at netflix. https://netflixtechblog.com/artwork-personalization-c589f074ad76, 2021. Accessed: January 2022. 3

[6] Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F Cohen. Bringing portraits to life. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017. 1

[7] David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. *arXiv preprint arXiv:2005.07727*, 2020. 3

[8] Amit H. Bermano, Rinon Gal, Yuval Alaluf, Ron Mokady, Yotam Nitzan, Omer Tov, Oren Patashnik, and Daniel

Cohen-Or. State-of-the-art in the architecture, methods and applications of stylegan. 2022. 3, 5

[9] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. *arXiv preprint arXiv:1707.05776*, 2017. 18, 19

[10] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 8

[11] Julie Cattiau. A communication tool for people with speech impairments. https://blog.google/outreach-initiatives/accessibility/project-relate/, 2021. Accessed: January 2022. 3

[12] Kelvin CK Chan, Xintao Wang, Xiangyu Xu, Jinwei Gu, and Chen Change Loy. Glean: Generative latent bank for large-factor image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14245–14254, 2021. 3

[13] Emily Denton, Ben Hutchinson, Margaret Mitchell, and Timnit Gebru. Detecting bias with generative counterfactual face attribute augmentation. *arXiv preprint arXiv:1906.06439*, 2019. 2

[14] Brian Dolhansky and Cristian Canton Ferrer. Eye in-painting with exemplar generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7902–7911, 2018. 3

[15] Michael S Floater. Generalized barycentric coordinates and applications. *Acta Numerica*, 24:161–214, 2015. 5

[16] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021. 3, 6

[17] Shiming Ge, Chenyu Li, Shengwei Zhao, and Dan Zeng. Occluded face recognition in the wild by identity-diversity inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10):3387–3397, 2020. 3

[18] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741, 1984. 3

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1, 3

[20] Jinjin Gu, Yujun Shen, and Bolei Zhou. Image processing using multi-code gan prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3012–3021, 2020. 3

[21] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. *arXiv preprint arXiv:2004.02546*, 2020. 3

[22] Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, 33(12):2341–2353, 2010. 3

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 8

[24] Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019. 3

[25] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. *arXiv:1710.10196*, 2017. 17

[26] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 3

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, pages 4401–4410, 2019. 1, 2, 4, 5, 8, 17

[28] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, pages 8110–8119, 2020. 4, 8

[29] Xiaoming Li, Wenyu Li, Dongwei Ren, Hongzhi Zhang, Meng Wang, and Wangmeng Zuo. Enhanced blind face restoration with multi-exemplar images and adaptive spatial feature fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2706–2715, 2020. 3

[30] Yijun Li, Richard Zhang, Jingwan Lu, and Eli Shechtman. Few-shot image generation with elastic weight consolidation. *arXiv preprint arXiv:2012.02780*, 2020. 3

[31] Xuan Luo, Xuaner Zhang, Paul Yoo, Ricardo Martin-Brualla, Jason Lawrence, and Steven M Seitz. Time-travel rephotography. *arXiv preprint arXiv:2012.12261*, 2020. 3, 5

[32] Sachit Menon, Alexandru Damian, Shijia Hu, Nikhil Ravi, and Cynthia Rudin. Pulse: Self-supervised photo upsampling via latent space exploration of generative models. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 2437–2445, 2020. 2, 3, 5

[33] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Freeze the discriminator: a simple baseline for fine-tuning GANs. *arXiv preprint arXiv:2002.10964*, 2020. 3

[34] Utkarsh Ojha, Yijun Li, Jingwan Lu, Alexei A Efros, Yong Jae Lee, Eli Shechtman, and Richard Zhang. Few-shot image generation via cross-domain correspondence. *arXiv preprint arXiv:2104.06820*, 2021. 2, 3, 8, 9

[35] Xingang Pan, Xiaohang Zhan, Bo Dai, Dahua Lin, Chen Change Loy, and Ping Luo. Exploiting deep generative prior for versatile image restoration and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[36] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. *arXiv preprint arXiv:2103.17249*, 2021. 3

[37] Dario Pavllo, David Grangier, and Michael Auli. Quaternet: A quaternion-based recurrent model for human motion. *arXiv preprint arXiv:1805.06485*, 2018. 6

[38] Justin NM Pinkney and Doron Adler. Resolution dependant GAN interpolation for controllable image synthesis between domains. *arXiv preprint arXiv:2010.05334*, 2020. 6

[39] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a StyleGAN encoder for image-to-image translation. In *Proc. IEEE/CVF CVPR*, pages 2287–2296, 2021. 3, 4

[40] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-shot adaptation of generative adversarial networks. *arXiv*, 2020. 3

[41] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *arXiv preprint arXiv:2106.05744*, 2021. 2, 3, 4, 6, 10, 12, 17

[42] Stefan Roth and Michael J Black. Fields of experts: A framework for learning image priors. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 860–867. IEEE, 2005. 3

[43] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterFaceGAN: interpreting the disentangled face representation learned by GANs. *arXiv preprint arXiv:2005.09635*, 2020. 3, 10, 12

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 7

[45] Nurit Spingarn-Eliezer, Ron Banner, and Tomer Michaeli. Gan" steerability" without optimization. *arXiv preprint arXiv:2012.05328*, 2020. 6

[46] Tommi Tervonen, Gert Valkenhoef, Nalan Basturk, and Douwe Postmus. Hit-and-run enables efficient weight generation for simulation-based multiple criteria decision analysis. *European Journal of Operational Research*, 224:552–559, 02 2013. 17

[47] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 3, 4, 6

[48] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization, 2017. 3

[49] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the GAN latent space. *arXiv preprint arXiv:2002.03754*, 2020. 3

[50] Neal Wadhwa, Rahul Garg, David E. Jacobs, Bryan E. Feldman, Nori Kanazawa, Robert Carroll, Yair Movshovitz-Attias, Jonathan T. Barron, Yael Pritch, and Marc Levoy. Synthetic depth-of-field with a single-camera mobile phone. *ACM Trans. Graph.*, 37(4), jul 2018. 10

[51] Kaili Wang, Jose Oramas, and Tinne Tuytelaars. Multiple exemplars-based hallucination for face super-resolution and editing. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[52] Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9168–9178, 2021. 3

[53] Zongze Wu, Dani Lischinski, and Eli Shechtman. StyleSpace analysis: Disentangled controls for StyleGAN image generation. *arXiv:2011.12799*, 2020. 1, 5

[54] Zongze Wu, Yotam Nitzan, Eli Shechtman, and Dani Lischinski. Stylealign: Analysis and applications of aligned stylegan models. *arXiv preprint arXiv:2110.11323*, 2021. 6

[55] Ceyuan Yang, Yujun Shen, Yinghao Xu, and Bolei Zhou. Data-efficient instance generation from instance discrimination. *arXiv preprint arXiv:2106.04566*, 2021. 3

[56] Tao Yang, Peiran Ren, Xuansong Xie, and Lei Zhang. Gan prior embedded network for blind face restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 672–681, 2021. 3, 10, 12

[57] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 3

[58] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4, 7, 8, 9, 18

[59] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2021. 9, 10

[60] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *arXiv preprint arXiv:2006.10738*, 2020. 2, 3, 8, 9, 10, 12

[61] Yajie Zhao, Weikai Chen, Jun Xing, Xiaoming Li, Zach Bessinger, Fuchang Liu, Wangmeng Zuo, and Ruigang Yang. Identity preserving face completion for large ocular region occlusion. *arXiv preprint arXiv:1807.08772*, 2018. 3

[62] Yijun Zhou and James Gregson. Whenet: Real-time fine-grained estimation for wide range head pose. *arXiv preprint arXiv:2005.10353*, 2020. 13

A. Dataset Curation

We collected a new dataset that contains sets of face images of celebrities. For each celebrity, we curated a set of high-quality images of faces of resolution 1024×1024. We used Amazon Mechanical Turk to filter images that do not match the identity of the wanted celebrity, images that contain occlusions of the face area, images with low resolution and watermarked images. Furthermore, to remove duplicated face images we calculated a similarity measure for each pair of images and filtered out images that were too close. Specifically, their deep features extracted from a classifier had cosine similarity greater than 0.9. Finally, the faces were aligned by the alignment process presented by Karras et al. [25] and separated into training sets and test sets for each celebrity. The number of images that were included in the test set and the training set of each celebrity presented in Table 5.

Table 5. The sizes of the training sets and test sets of our dataset.

Celebrity	Training set size	Test set size
Adele	109	8
Angela Merkel	145	10
Barack Obama	192	13
Dwayne Johnson	97	11
Emilia Clarke	258	11
Jeff Bezos	114	3
Joe Biden	206	13
Kamala Harris	110	7
Lady Gaga	133	6
Michelle Obama	279	9
Oprah Winfrey	135	9
Taylor Swift	158	11
Xi Jinping	92	15

B. Additional Details

B.1. Sampling From $\mathcal{P}_0$

Commonly, sampling from generative models is trivial as the probability density over the latent space is known. Conversely, $\mathcal{P}_0$ is defined as the area enclosed by a convex hull. We therefore require a sampling strategy to sample from $\mathcal{P}_0$. We propose a simple protocol where we uniformly sample a random vector from $\mathcal{A}_0$, in which there are three non-zero entries. This is trivially done by sampling three anchors and three scalars from $[0, 1)$ which are then normalized. From that vector, we obtain the latent code in $\mathcal{P}_0$. We note that there exists a large body of works on sampling from convex bodies [46], which may produce superior results. However, this investigation is out of the scope of this paper.

B.2. Implementation Details

All projection experiments, including inversion, inpainting and super-resolution inherit the setting and hyperparameters from StyleGAN2 projection code [27]. This includes the optimizer, number of steps, learning rate schedulers, etc. Similarly, the tuning process inherits its hyperparameters from PTI [41]. For both tuning our model and projecting into it, we observe no need for early-stopping. On the contrary, all projections to DiffAugment were stopped after 200 iterations. We observe that more iterations lead to significant degradation in their results.

Hyperparameters values used in experiments – $s = 100$ in Eqn. (7), $\lambda_{L_2} = 1$ in Eqn. (1), $\lambda_{d-reg} = 10$ and applied only on the first 12 layers in Eqn. (9).

Tuning our model takes roughly 48 seconds per anchor on a single V100 GPU. This translates to roughly 80 minutes for 100 anchors.

Previous methods performing optimization in latent space [1] have found it beneficial to initialize the optimization to the mean code in latent space – $\bar{w}$. Similarly, we initialize our optimization to the center of $\mathcal{P}$. This corresponds to initializing $\boldsymbol{\alpha}$ to a vector whose all components are equal.

B.3. Anchors' Linear Independence

In several occasions throughout the paper, we have implicitly assumed that the obtained anchors are linearly independent. We first note that in practice, this was the case in all of our experiments. However, this might not always be the case, specifically as N increases and becomes close to k or surpasses it. Linearly dependent anchors have no strong impact on our method. One can drop the anchors that are internal to the convex hull and obtain a linearly independent set spanning the same $\mathcal{P}$. Doing so is mostly useful as a means to reduce the dimension of α-space and for projecting to the span of the anchors. Synthesis and projection may seamlessly operate with a dependent set of anchors.

C. Additional Experiments

C.1. Effect of β on Projection

Throughout our experiments, we have demonstrated that latent codes on the $\mathcal{P}_\beta$ manifold follow a personalized, high-quality prior. We additionally demonstrated that other latent codes may not follow the prior (see Figures 3, 5 and 6).

Specifically, in Section 4.1 and Figure 5, we have demonstrated that the prior is local by observing gradual vanishment of personalization from synthesized images along traversals outwards from the manifold. Greater β values are required for containing further traversals, and are hence directly associated with vanishment of personalization.

In this section, we complement this experiment, by evaluating the effect of β on projecting given images to $\mathcal{P}_\beta^+$. We use the projection method described in Section 3.3, to solve the tasks of inversion, inpainting and super-resolution. Visual results are provided in Figure 15.

The projection to $\mathcal{P}_0^+$ yields a highly characteristic image of the person. However, it is also strongly conservative, *i.e.* relating to a common and simple appearance of the person – almost frontal, neutral or smiling expression, simple illumination conditions and no accessories. Therefore, across applications, typical appearances are more accurately reconstructed than atypical appearances.

One can also observe that, in most cases, increasing β corresponds to greater expressivity, indicated by better fidelity. This is most evident for atypical images. We find that for such images, the optimization in fact converges to latent codes outside the previous smaller dilation. On the other hand, we observe that, "typical" images that were properly reconstructed with the smaller dilation, converge to a similar dilation and do not leverage the newly enlarged one.

This behavior is intuitive due to the optimization's dynamic. As a reminder – β values serve as upper-bounds in Eqn. (7), and projection is initialized to the center of $\mathcal{P}_0^+$. Therefore, the optimization tends to converge at a smaller dilation when possible.

While increasing β improves fidelity, we also observe it is associated with artifacts and slight drifts to identity. Once more this is true mostly for projections that resort to a greater effective dilation.

We conclude that β continuously controls a tradeoff between the prior and expressiveness. The actual dilation the optimization converges to, depends on how conservative the input image is.

Choosing a value for β depends on user's preference for the balance between fidelity to quality and personalization. In our experiments, we use the minimal β that was able to obtain fair fidelity. We note that this simple guideline depends on the application. For example, when solving inpainting, only a small portion of the generated image is used by blending it to the input image. Therefore, obtaining fair fidelity is possible with smaller β, compared to the one needed for fair fidelity in inversion and super-resolution.

C.2. Nearest-Neighbor Experiments

We last demonstrate that the outputs of our method are not merely duplicates of the reference set. In Figure 16, we display inpainting, super-resolution and synthesis results along with their LPIPS [58] nearest-neighbor from the reference set. As can be seen, our result resembles the nearest-neighbor which is expected but is never a simple duplicate.

C.3. Predetermined vs Trained Anchors

When adapting the pre-trained generator G_d, we minimize the reconstruction loss on a set of images $\{x_i\}_{i=1}^N$ with their corresponding predetermined anchors $\{w_i\}_{i=1}^N$. We might ask, *do we need to predetermine the anchors?*

If we instead let anchors train together with the generator, the resulting process is similar to generative latent optimization (GLO) [9]. We next evaluate the effect of replacing our tuning approach with GLO. Figure 17 presents random images synthesized with the obtained GLO generator for three individuals - Adele, Kamala Harris and Joe Biden. As can be seen, the synthesized images are less realistic and exhibit considerable artifacts and blurriness as well as distortions in key-facial characteristics of the person. To conclude, optimizing the anchors along with generator causes inferior results to those obtained with predetermined anchors as demonstrated in Figure 8.

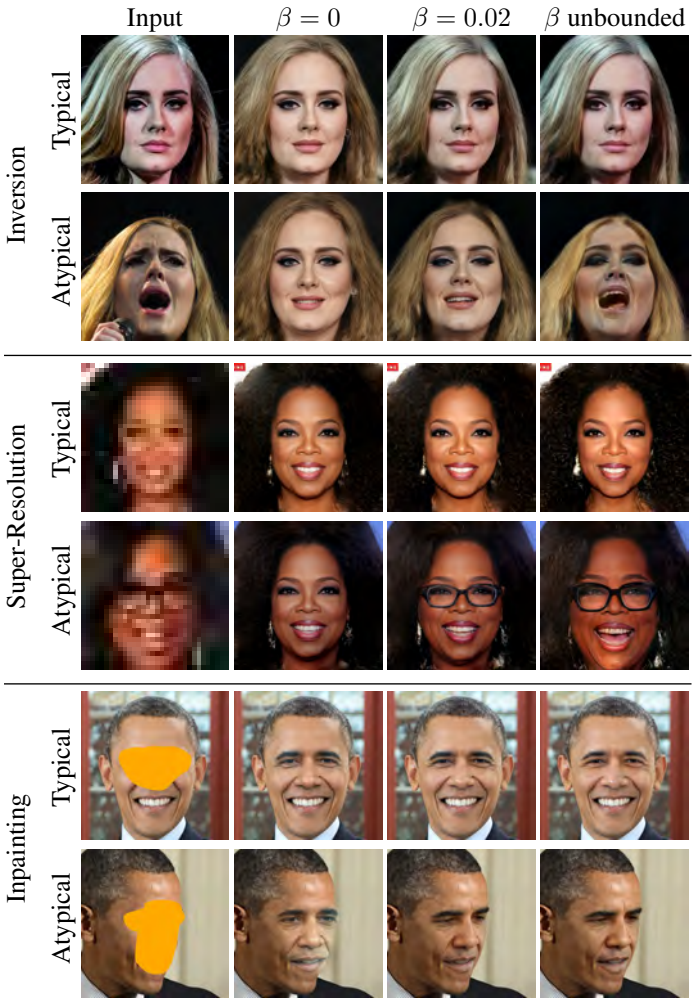


Figure 15. Effect of β on projection for inversion, super-resolution and inpainting. The effect is demonstrated for both typical and atypical images for each person and application. It can be seen that β controls a tradeoff between the personalized prior and expressiveness. Results produced with $\beta = 0$ are highly characteristic and conservative. These results might be sufficient for some cases (e.g. rows 3, 5) but have poor fidelity in many other. Increasing the allowed β, allows greater expressivity and thus better fidelity, but at the cost of weakening the guidance of the prior. At the extreme, β is not explicitly bounded (Eqn. (7) not used). These results might be viable in some cases (e.g. rows 1, 3, 6) but may also produce results containing artifacts (e.g. rows 2, 4) or slight drift in identity (e.g. row 5). We find it beneficial in all applications to bound β to some positive small value, in order to balance the trade-off.

(a) Synthesis (b) Inpainting (c) Super-Resolution

Figure 16. Nearest-Neighbor experiments. We present images generated by our models for the tasks of synthesis, inpainting and super-resolution along side their respective LPIPS nearest-neighbors (marked NN). For image enhancement results, the inputs are also displayed. As can be seen, in no case are the results obtained by our method merely duplicates of their nearest neighbor in the reference set.

Figure 17. Random images synthesized by adapting the generator using GLO [9]. The synthesized images are blurry, less realistic and exhibit subtle distortions in the key-facial characteristics of the person, in contrast to our results in Figure 8.