Controllable GAN Synthesis Using Non-Rigid Structure-from-Motion

René Haas Stella Graßhof Sami S. Brandt
IT University of Copenhagen, Denmark
{renha,stgr,sambr}@itu.dk

Abstract

In this paper, we present an approach for combining non-rigid structure-from-motion (NRSfM) with deep generative models, and propose an efficient framework for discovering trajectories in the latent space of 2D GANs corresponding to changes in 3D geometry. Our approach uses recent advances in NRSfM and enables editing of the camera and non-rigid shape information associated with the latent codes without needing to retrain the generator. This formulation provides an implicit dense 3D reconstruction as it enables the image synthesis of novel shapes from arbitrary view angles and non-rigid structure. The method is built upon a sparse backbone, where a neural regressor is first trained to regress parameters describing the cameras and sparse non-rigid structure directly from the latent codes. The latent trajectories associated with changes in the camera and structure parameters are then identified by estimating the local inverse of the regressor in the neighborhood of a given latent code. The experiments show that our approach provides a versatile, systematic way to model, analyze, and edit the geometry and non-rigid structures of faces.

1. Introduction

In recent years, Generative Adversarial Networks (GANs) [15] have seen rapid improvements in image quality as well as training stability. GANs have achieved remarkable results in tasks such as image synthesis [23–27], image-to-image translation [11, 12, 36], semantic editing [1,2,21,33,39,43,47] as well as regression tasks [32]. Especially the StyleGAN [24–27] family of models show state-of-the-art results in unconditional synthesis human faces images. However, the standard StyleGAN architecture provides no way to directly control semantics like the pose and expression of the generated images. This has led to a large interest in finding semantic directions in the latent space of StyleGAN which controls specific semantic attributes such as pose, expression, hairstyle, illumination, etc.

The non-rigid structure-from-motion (NRSfM) problem is a difficult, under-constrained problem with a long history in computer vision. NRSfM aims at obtaining the three-dimensional reconstruction of a scene with dynamical deformable structures from a sequence of 2D correspondences. Given a set of 2D correspondences, the standard assumption is that the deformable 3D shape is a linear combination of basis shapes; the camera information, describing how the 3D structure is projected onto the image plane, also needs to be recovered. In this work, we incorporate a sparse 3D model based on NRSfM into a generative model like StyleGAN. This is interesting for two reasons: first, this allows us to find trajectories in the latent space corresponding to well-defined semantic attributes corresponding to the camera geometry and non-rigid structure. Second, using a generative model in conjunction with NRSfM provides a

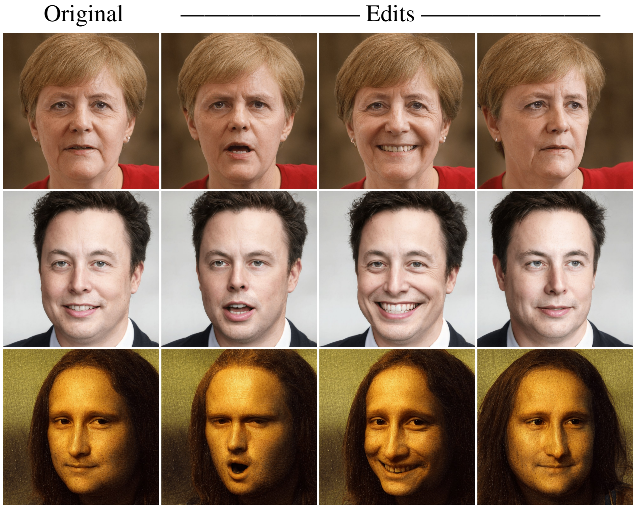


Figure 1. **Semantic editing of real image.** Our method parameterizes the latent space of StyleGAN in terms of camera and shape parameters. This allows for editing of rotation, translation, and non-rigid shape deformation of the synthesized images. Coupled with a strong latent encoder, like e4e [45] or HyperStyle [5], our method allows for semantic editing of real images. Here we show two non-rigid changes corresponding to facial expressions (2nd and 3rd column) as well as a rigid edit corresponding to camera orientation (4th column).

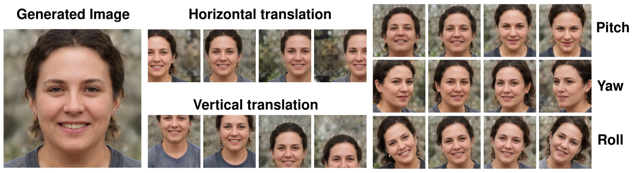


Figure 2. **Rigid edits to rotation and translation.** Our method discovers trajectories in latent space corresponding to arbitrary rotations and translation.

way to obtain an *implicit* dense 3D reconstruction by using only the sparse 2D inputs. By this, we refer to the fact that we are able to view the dense 2D face from an arbitrary 3D orientation, as if we had an explicit dense 3D reconstruction available. In other words, our approach enables dense image synthesis of novel shapes from arbitrary view angles and non-rigid deformation without the need for an explicit dense 3D reconstruction.

In Fig. 1, we demonstrate semantic editing of real images by using our method in conjunction with a recent method for GAN inversion [45]. In Fig. 2 we show latent trajectories corresponding to changes in the rigid camera parameters such as rotation and translations. Note that such edits are only possible if the generator has been trained on a data set that contains such variations, *i.e*., of translation and roll rotation. In other words, we need an unaligned data set, like FFHQU [25].

Our method utilizes a sparse backbone that is a 3D model based on the approach for NRSfM given in [8, 16]. The 3D model is constructed using solely 2D landmarks extracted from synthetic face images generated by StyleGAN, thus our approach requires no 3D supervision.

In this approach, we first factorize the measurement matrix, consisting of corresponding 2D landmark points, into a rigid and non-rigid part each composed of camera and 3D shape information respectively. Any arbitrary 3D shape can then be represented as the sum of a rigid basis shape and a linear combination of rank-one non-rigid basis shapes. Our approach provides a way to recover a set of expansion coefficients that contains all the information about the 3D reconstruction of the extracted 2D face landmarks. Additionally, for each set of 2D landmarks, we recover a projection matrix, describing the camera information for projecting the 3D shapes onto the image plane as well as information about the orientation of the recovered 3D structure.

We then proceed to connect the information recovered from the sparse 2D landmarks to the latent space of StyleGAN by training a regressor in the form of a multilayer perceptron (MLP) network to regress the shape and camera information directly from the latent codes. By estimating the local inverse of the regressor at a given latent code, we can identify trajectories in latent space corresponding to changes in camera or non-rigid geometry, while preserving other attributes of the generated image, like identity, texture, and illumination. We show that the regressor network can be used for semantic editing of latent codes, either by using the first-order Taylor expansion of the trained network to define linear directions in latent space or by using the prediction of the network as a loss term for a gradient-based optimization algorithm.

As noted in [46], performing semantic editing in StyleGAN using only 2D landmarks is a very challenging problem since the 2D coordinates are extremely localized compared to more global attributes like age or gender.

In summary, we propose an editing framework that relies solely on sparse 2D landmarks. From the landmarks, we use NRSfM to extract camera and shape parameters describing the underlying 3D geometry. We train a regressor to predict these parameters directly from the latent codes and show how the regressor naturally enables editing of the camera and non-rigid geometry of the generated images.

The main contributions of this paper are the following.

- We propose a framework that incorporates the NRSfM problem into the latent space of generative models.
- Based on NRSfM we suggest a framework to get artistic control over images synthesized by StyleGAN.
- We show how our approach can model the camera, pose, and non-rigid structure of the synthesized images, without an explicit dense 3D reconstruction.
- We propose a general method for enabling 3D awareness in 2D GANs without requiring any retraining or changes to the generator architecture.
- We propose a regularization technique that preserves the identity of the synthesized faces during the edits.

2. Related Work

StyleGAN. The StyleGAN [24–27] generator is inspired by the style transfer literature [14, 20] and consists of a *mapping network* f which maps a latent vector $\mathbf{z} \in \mathcal{Z}$, sampled from the standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ in order to obtain an intermediate representation $\mathbf{w} \in \mathcal{W}$. The latent space $\mathcal{W}$ is more disentangled than $\mathcal{Z}$ [26]. To synthesize an image, the latent code $\mathbf{w}$ is copied and fed to each synthesis block of the *synthesis network* G which produces the final image. Instead of feeding the same vector to each of the synthesis blocks, if the vectors are allowed to differ, the resulting space is typically denoted as $\mathcal{W}+$. It has been shown that using $\mathcal{W}+$ space can lead to lower reconstruction loss when performing GAN inversion [35, 50], however at the cost of lower editability [45] of the resultant latent codes.

Semantic Editing. Several methods have been proposed to enable semantic edits of the images produced by StyleGAN. InterFaceGAN [38, 39] enables editing of binary semantic attributes like left/right pose, gender, presence or absence of smile, etc. Here, a set of latent codes are first sampled and the images are annotated using pre-trained binary classifiers. Following the annotation step, a support vector machine was fitted on the labeled data for each binary semantic attribute. The normal vector for the supporting hyperplane then defines the semantic direction in latent space. Another approach for semantic editing is GANSpace [21] which proposes to use PCA on sampled latent codes to find semantic directions in an unsupervised fashion. Another related approach also factorizes the weights of the trained generator [40, 42] rather than the latent codes. Both methods then change the semantics of the generated images by perturbing latent codes in the direction of the found semantic directions. Additionally, [2] uses normalizing flows for attribute-conditioned semantic editing and explores both linear and non-linear trajectories in latent space. Another related approach, StyleRIG [43] proposes semantic editing in StyleGAN using 3D morphable models [7]. Recently it was proposed to regard the space of channel-wise style parameters after the learned affine transformation in each block in the StyleGAN synthesis network as a separate latent space, complementing the previously mentioned $\mathcal{Z}$, $\mathcal{W}$ and $\mathcal{W}+$ spaces. This latent space was named StyleSpace and denoted as $\mathcal{S}$ [47]. It has been shown that $\mathcal{S}$ space has superior disentanglement properties, especially in StyleGAN3 [4, 25], compared to $\mathcal{W}$ space thus enabling fine-grained and highly localized edits, like the closing of the eyes or changes to hair color [47].

Inversion. For purposes involving the editing of real images, it is necessary to find a good latent representation. That is, we need to find a latent code that, when passed to the generator, reconstructs the target image. This problem is known as GAN inversion. Techniques for GAN inversion have either used optimization-based approaches, where the latent code is directly optimized in order to reconstruct the target image [1, 27, 35] or encoder-based approaches, where a target image is directly mapped into the latent space [3, 34, 36] or hybrid approaches [6, 50].

Recent work [45] suggests that there is a trade-off between distortion and editability when selecting which latent space to project a given target image into. Projecting images into the extended $\mathcal{W}+$ space typically leads to higher reconstruction quality [35], *i.e.*, produces a generated image which is more similar to the target image. However, latent codes in $\mathcal{W}+$ are generally less suitable for semantic editing than latent codes in the native $\mathcal{W}$ space.

The e4e encoder proposed in [45] seeks to find a good trade-off between reconstruction and editability by projecting images into $\mathcal{W}+$ but constraining the latent codes to be close to $\mathcal{W}$. Recently [37] shows that real images can be embedded into $\mathcal{W}$ space by fine-tuning the trained generator around the target image, thus circumventing the need for projecting into $\mathcal{W}+$ space. In [3], a combination of the iterative and encoder-based methods is proposed. Here the encoder predicts the residual with respect to the current estimate of the latent code and thus is able to refine the latent code using only a few forward passes of the encoder in a process referred to as iterative refinement. Recently, [5] proposed to unite the ideas of fine-tuning the generator from [37] with the iterative refinement from [3] by introducing a hypernetwork which predicts how the parameters of the generator should be changed in order to faithfully embed a given real image into the native, and more editable, $\mathcal{W}$ space.

Explicitly 3D aware GANs. Several works have investigated incorporating explicit 3D understanding into GANs [17, 31, 48]. Compared to these, our approach can be used to control the 3D structure in existing 2D GANs without the need for adaptation of the generator architecture nor does our approach require any retraining.

NRSfM. Structure-from-motion (SfM) deals with the problem of inferring the scene geometry and camera information from image sequences. In [44], an orthographic camera model was assumed to infer rigid shape and motion by a factorization of the measurement matrix. In [10], this problem was formulated to include non-rigid deformations by assuming that a shape is a linear combination of 3D basis shapes, hence proposing an approach for non-rigid structure-from-motion (NRSfM). Various works have followed up on this approach over the years this is still an area of active research [22].

Recently, there have been attempts to solve the NRSfM problem by employing neural networks. However, most require a large training data set [29], 3D supervision, or an assumption of an orthographic camera model [29,41]. Specifically, [29] formulates the NRSfM problem as a multi-layer block sparse dictionary learning problem converted into a deep neural network. In neural NRSfM [41], the authors rely on dense 2D point tracks to recover dense 3D representations, and train an auto-decoder-based model with subspace constraints in the Fourier domain. Our method differs from these works in several aspects, because (1) it relies only on sparse 2D points, (2) it does not rely on a block structure, and (3) it assumes an affine camera model. This makes our approach direct, lightweight, fast, and efficient.

3. Method

Let $I = G(\mathbf{w})$ be an image generated by the StyleGAN generator by the latent code $\mathbf{w}$. Our goal is to locally parameterize the manifold of latent codes, in the neighborhood of a fixed latent code $\mathbf{w}_0$, by an *attribute vector* $\mathbf{q}$ so that

$$\mathbf{w} = \Omega_{\mathbf{w}_0}(\mathbf{q}), \tag{1}$$

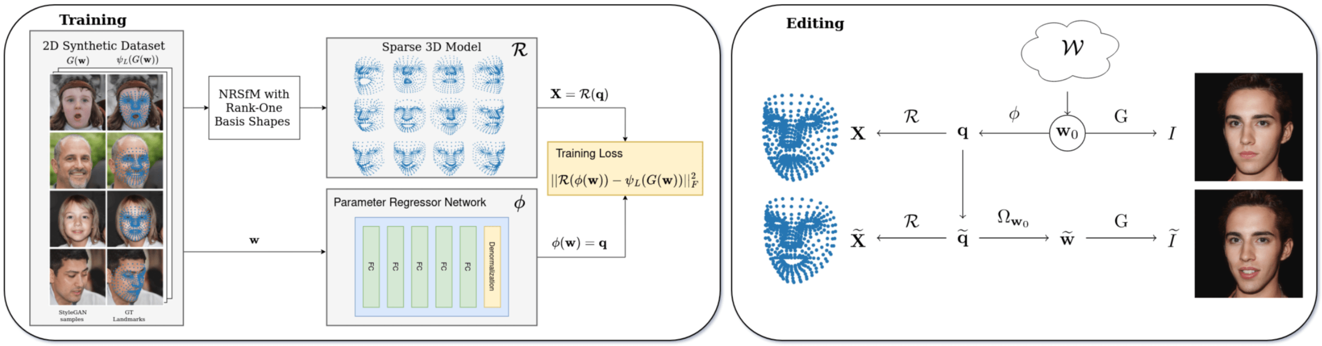


Figure 3. **Overview of our method.** We first create a sparse 3D model $\mathcal{R}$ of facial landmarks from a data set of 2D landmarks $\mathbf{X}$ using NRSfM. The 3D model is parameterized by an attribute vector $\mathbf{q}$ which contains information about the camera, rotation, and non-rigid 3D structure. We then train a regressor ϕ to predict the parameters $\mathbf{q}$ directly from latent codes $\mathbf{w}$. Once the regressor is trained, it can be used for semantic editing. Given a latent code $\mathbf{w}_0$ with corresponding attribute vector $\mathbf{q}_0$ we can define a different, target attribute vector $\widetilde{\mathbf{q}}$ and transfer it onto $\mathbf{w}_0$ using the transformation $\Omega_{\mathbf{w}_0}$ which depends on the regressor ϕ.

where $\mathbf{q}$ describes the *pose*, *shape*, and *camera* information of the generated image. This formulation facilitates the transfer of the target attributes $\mathbf{q}$ onto the latent code $\mathbf{w}_0$ to obtain an edited code $\mathbf{w}$ where only the target attributes have changed in the image, while preserving all other attributes such as identity, texture, and illumination.

Our method is composed of three distinct elements. (1) The *sparse back-bone* relies on a pre-trained landmark extractor ψ_L, which extracts the 2D landmarks $\mathbf{X} = \psi_L(I)$, from a generated image I coupled with a closed-form parameterization for the 2D landmarks as $\mathbf{X} = \mathcal{R}(\mathbf{q})$, where $\mathcal{R}$ maps the 3D shape defined by the attribute vector $\mathbf{q}$ onto the image plane. (2) The *attribute regressor* ϕ predicts the attribute vector $\mathbf{q} = \phi(\mathbf{w})$ from the latent code $\mathbf{w}$, where the regressor is trained by minimizing the squared distance between the ground truth landmarks $\mathbf{X} = (\phi_L \circ G)(\mathbf{w})$ and predicted landmarks $\widehat{\mathbf{X}} = (\mathcal{R} \circ \phi)(\mathbf{w})$. (3) The *regression inversion* constructs the local inverse of the regressor ϕ around the latent code $\mathbf{w}_0$, *i.e.*, finds the local parameterization of the latent space so that $\mathbf{w} = \Omega_{\mathbf{w}_0}(\mathbf{q})$, where $\phi(\mathbf{w}_0) = \mathbf{q}_0$. In Fig. 3 we provide a graphical overview of our approach.

The remaining part of this section is organized as follows. In Section 3.1 we introduce the landmark parameterization $\mathcal{R}(\mathbf{q})$ and detail how the 3D basis shapes can be recovered from a data set of sparse 2D landmarks. The training of the attribute network ϕ is discussed in Section 3.2 and finally, in Section 3.3 we show how the regressor ϕ is used to facilitate highly interpretable semantic editing.

3.1. Rank-one model

The rank-one approach for non-rigid structure-from-motion, proposed in [8, 9, 16], is an affine camera model for non-rigid structure-from-motion which is able to recover 3D structure from sparse 2D correspondences using rank-one basis shapes. In this paper, we frame the model as a parameterization of the space of possible 2D shapes in terms of camera, rotation, translation, and shape parameters. We propose to write the model of [8, 9] in closed-form as

$$\mathcal{R}(\mathbf{q}) = \underbrace{\mathbf{K}[\mathbf{I}_2|\mathbf{0}]\mathbf{R}(\boldsymbol{\theta})}_{\mathbf{M}} \left[\mathbf{B}_0 + \sum_{k=1}^{K} \alpha_k \mathbf{B}_k \right] + \mathbf{t} \otimes \mathbf{1}_L^{\mathrm{T}}, \quad (2)$$

where $\mathbf{K} \in \mathbb{R}^{2\times 2}$ an upper triangular matrix, containing the camera parameters $\mathbf{k} = (k_{11}, k_{12}, k_{22})$. The rotation matrix $\mathbf{R} \in \mathbb{R}^{3\times 3}$ is parameterized in terms of the Euler angles $\boldsymbol{\theta} = (\theta_x, \theta_y, \theta_z)$. The rigid basis shape $\mathbf{B}_0$ describes the average 3D reconstruction while the non-rigid basis shapes $\mathbf{B}_k$ for $k > 0$ describe the non-rigid variation from the rigid basis shape. The expansion coefficients $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_K)$ determine the strength of the contribution of each of the non-rigid basis shapes $\mathbf{B}_k$. Finally, the translation vector $\mathbf{t}$ determines the offset from the origin. In (2), $\otimes$ denotes the Kronecker product, $\mathbf{1}_L \in \mathbb{R}^L$ is a vector of ones, thus $\mathbf{t}\otimes\mathbf{1}_L^{\mathrm{T}} \in \mathbb{R}^{2\times L}$ yields a matrix where $\mathbf{t} \in \mathbb{R}^2$ is repeated L-times column-wise. To summarize, with (2) any 2D shape $\mathbf{X}$ can be parameterized in terms of an attribute vector $\mathbf{q}$ as $\mathbf{X} = \mathcal{R}(\mathbf{q})$ where the attribute vector contains the camera, rotation, shape, and translation parameters as $\mathbf{q} = (\mathbf{k}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{t})$.

In the next section, we see how the rigid basis shape $\mathbf{B}_0$ and non-rigid basis shapes $\mathbf{B}_k$, $k > 0$, can be recovered given a data set of corresponding 2D landmark points.

3.1.1 Non-rigid Factorization.

Given N 2D shapes $\mathbf{X}_n \in \mathbb{R}^{2\times L}$, we stack them into a measurement matrix $\mathcal{X} \in \mathbb{R}^{2N\times L}$. Our aim is to factorize $\mathcal{X}$ into a rigid $\mathbf{X}_0$ and non-rigid $\delta\mathbf{X}$ part such that

$$\mathcal{X} = \mathcal{X}_0 + \delta\mathcal{X} = \mathbf{M}_0\mathbf{B}_0 + \delta\mathbf{M}\delta\mathbf{B}. \quad (3)$$

To recover the rigid basis shape $\mathbf{B}_0$ from (2) we first calculate the singular value decomposition (SVD) of the measurement matrix as $\mathcal{X} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\mathrm{T}}$. The rigid part $\mathcal{X}_0$ is then constructed by selecting the three dominant singular vectors such that

$$\mathcal{X}_0 = \mathbf{U}_0\mathbf{\Lambda}_0\mathbf{V}_0^{\mathrm{T}} = \mathbf{M}_0\mathbf{B}_0 \quad \text{with} \tag{4}$$

$$\mathbf{M}_0 = \mathbf{U}_0\mathbf{\Lambda}_0 \in \mathbb{R}^{2N\times 3}, \quad \mathbf{B}_0 = \mathbf{V}_0^{\mathrm{T}} \in \mathbb{R}^{3\times L}. \tag{5}$$

The matrix $\mathbf{M}_0$ contains the N affine projection matrices $\mathbf{M}_n$, associated with each shape in the data set, which are stacked on top of each other in $\mathbf{M}_0$.

To recover the non-rigid basis shapes $\mathbf{B}_k$, we subtract the rigid part from the measurement matrix, *i.e.*, $\delta\mathcal{X} = \mathcal{X} - \mathcal{X}_0$, and calculate the SVD of the remaining part as

$$\delta\mathcal{X} = \delta\mathbf{U}\delta\mathbf{\Lambda}\delta\mathbf{V}^{T} = \delta\mathbf{M}\delta\mathbf{B}. \tag{6}$$

In the following, we use $\delta\mathbf{B} = \delta\mathbf{V}^{\mathrm{T}} \in \mathbb{R}^{L\times L}$ to construct the non-rigid basis shapes as $\mathbf{B}_k = \mathbf{d}_k\mathbf{b}_k^{\mathrm{T}}$, where $\mathbf{b}_k^{\mathrm{T}}$ is the kth row of $\delta\mathbf{B}$, and $\mathbf{d}_k$ is a 3×1 unit vector which will be determined by gradient-based optimization. Now our goal is to recover $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_K] \in \mathbb{R}^{3\times K}$ which defines the non-rigid basis shapes. In [8, 9, 16], $\mathbf{D}$ was recovered by an alternating least squares optimization scheme by exploiting the orthonormality of the non-rigid basis shapes. Here we use gradient-based optimization instead. For this purpose, it is convenient to write the factorization of the measurement matrix $\mathcal{X}$ as

$$\mathcal{X} = \mathbf{M}_0\mathbf{B}_0 + \mathbf{M}^{\alpha}\mathbf{B}, \tag{7}$$

where

$$\begin{aligned}\mathbf{M}^{\alpha} &= (\boldsymbol{\alpha} \otimes \mathbf{1}_{2\times 3}) \odot (\mathbf{1}_K \otimes \mathbf{M}_0) \\ &= \begin{bmatrix} \alpha_{11}\mathbf{M}_1 & \alpha_{12}\mathbf{M}_1 & \cdots & \alpha_{1K}\mathbf{M}_1 \\ \alpha_{21}\mathbf{M}_2 & \alpha_{22}\mathbf{M}_2 & \cdots & \alpha_{2K}\mathbf{M}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{N1}\mathbf{M}_N & \alpha_{N2}\mathbf{M}_N & \cdots & \alpha_{NK}\mathbf{M}_N \end{bmatrix},\end{aligned} \tag{8}$$

where $\odot$ is the Hadamard product and

$$\mathbf{B} = \mathrm{diag}(\mathrm{vec}(\mathbf{D}))(\mathbf{I}_K \otimes \mathbf{1}_3)\delta\mathbf{B} = \begin{bmatrix} \mathbf{d}_1\mathbf{b}_1^{\mathrm{T}} \\ \mathbf{d}_2\mathbf{b}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{d}_K\mathbf{b}_K^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_K \end{bmatrix}. \tag{9}$$

Then we can jointly find $\mathbf{D}$ and $\boldsymbol{\alpha}$ by minimizing

$$\min_{\mathbf{D},\boldsymbol{\alpha}} ||\widehat{\mathcal{X}}(\mathbf{D},\boldsymbol{\alpha}) - \mathcal{X}||_F^2 + \lambda\sum_{k=1}^{K}(\mathbf{d}_k^{\mathrm{T}}\mathbf{d}_k - 1)^2, \ \lambda \in \mathbb{R}^+, \tag{10}$$

by gradient descent. Once we have found the $\mathbf{D}$ and $\boldsymbol{\alpha}$ which minimizes (10), the non-rigid basis shapes can be constructed using (9). The found basis shapes $\mathbf{B}_i$ completely specify the parameterization in (2).

The parameterization of a new unseen set of landmarks $\mathbf{X}_{\text{new}}$ can be obtained as

$$\mathbf{q}^* = \arg\min_{\mathbf{q}} ||\mathcal{R}(\mathbf{q}) - \mathbf{X}_{\text{new}}||_F^2. \tag{11}$$

3.2. Connection to the latent space

Having found the parameterization $\mathcal{R}$ in (2), we train a MLP network ϕ to regress the parameters $\mathbf{q}$ directly from the latent codes $\mathbf{w}$ such that $\phi(\mathbf{w}) = \widehat{\mathbf{q}}$. Predicting $\mathbf{q}$ is equivalent to predicting the landmarks of the generated images as $\mathcal{R}(\phi(\mathbf{w})) = \widehat{\mathbf{X}}$. We train the network ϕ to minimize the objective function

$$\mathcal{L}(\mathbf{w}) = \|\mathcal{R}(\phi(\mathbf{w})) - \psi_L(G(\mathbf{w}))\|_F^2, \tag{12}$$

where ψ_L is some pre-trained landmark extractor.

3.3. Semantic Editing

In the following, we provide an analytic as well as a gradient-based approach for locally inverting the trained network ϕ, to control the pose and non-rigid shape of images generated by StyleGAN. For the *analytic approach*, the first order Taylor expansion of ϕ around $\mathbf{w}_0$ yields

$$\phi(\mathbf{w}) = \phi(\mathbf{w}_0) + \mathbf{J}|_{\mathbf{w}=\mathbf{w}_0}(\mathbf{w} - \mathbf{w}_0), \tag{13}$$

where $\mathbf{J}|_{\mathbf{w}=\mathbf{w}_0}$ is the Jacobian of ϕ evaluated at $\mathbf{w}_0$. Now since $\phi(\mathbf{w}_0) = \mathbf{q}_0$ we can rewrite this as

$$\mathbf{w} = \mathbf{w}_0 + \mathbf{J}^{\dagger}(\mathbf{q} - \mathbf{q}_0), \tag{14}$$

where $\mathbf{J}^{\dagger}$ is the Moore-Penrose pseudo-inverse of $\mathbf{J}|_{\mathbf{w}=\mathbf{w}_0}$. This allows us to edit a latent code $\mathbf{w}_0$ with associated 2D landmarks $\mathbf{X}_0$ parameterized by $\mathbf{q}_0$ as $\mathbf{X}_0 = \mathcal{R}(\mathbf{q}_0)$ in such a way as to obtain a new latent code $\mathbf{w}$ with a corresponding set of landmarks parameterized by $\mathbf{q}$.

The analytic method described in (14) requires evaluating $\mathbf{J}$ at $\mathbf{w}_0$ and defines a linear path in latent space. As an alternative to (14) we propose a *gradient-based approach* where we directly minimize the difference between the network prediction $\phi(\mathbf{w})$ and a target attribute vector $\mathbf{q}_{\text{target}}$ via

$$\min_{\mathbf{w}} \|\phi(\mathbf{w}) - \mathbf{q}_{\text{target}}\|^2 + \lambda\mathcal{D}(G(\mathbf{w}), G(\mathbf{w}_0)), \tag{15}$$

where $\mathcal{D}(\cdot,\cdot)$ is an image similarity metric such as Learned Perceptual Image Patch Similarity (LPIPS) [49] or Arcface [13], which we employ for regularization purposes. The gradient-based editing is analogous to what is proposed in [46]. However, here we allow for the passing of gradients

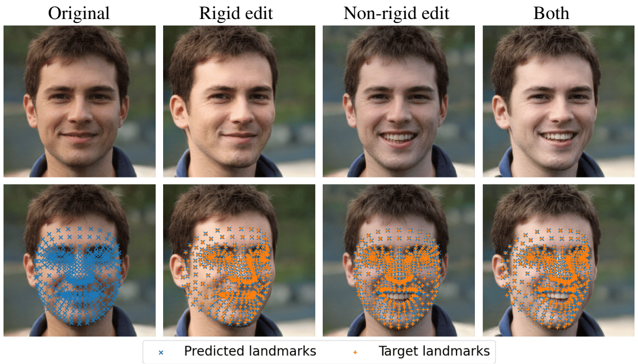


Figure 4. **Rigid and non-rigid edits.** Our approach disentangles rigid edits (rotation) from non-rigid edits (facial expression). We observe that the predicted landmarks agree well with the target landmarks for both types of edits.

through the generator G in order to calculate the identity loss in (15).

In Fig. 4 we visualize the landmarks predicted by our regressor from the latent code as $\mathcal{R}(\phi(\mathbf{w}))$ in blue. Additionally, we showcase semantic editing by changing the latent code $\mathbf{w}$ towards a set of target landmarks $\mathcal{R}(\mathbf{q}_{\text{target}})$ in orange. We show a rigid edit of camera rotation, by changing $\boldsymbol{\theta}$ and a non-rigid edit to facial expression by changing $\boldsymbol{\alpha}$, as well a combination.

4. Experiments

4.1. Implementation Details

We used the StyleGAN2 [27] networks pre-trained on FFHQ [26] as well as StyleGAN3 [25] pre-trained on FFHQU [25]. FFHQ consists of $70K$ face images from flicker and FFHQU is the unaligned version. To construct the model $\mathcal{R}$ in (2) we first sampled $N = 5 \times 10^4$ synthetic images and from each extracted $L = 68$ landmark points with Dlib [28] and $L = 468$ using MediaPipe [30], which were then normalized to the interval $[0, 1]$. In each of the following experiments, we have set the number of non-rigid basis shapes to $K = 12$. Further, we rotated the basis shapes to face the camera when $\boldsymbol{\theta} = \mathbf{0}$ in (2) in order to stabilize the training of the regressor. We trained the regressor, to predict the mean-centered output features $\widehat{\mathbf{q}}$ for each of the N samples. We used the Adam optimizer, 3 hidden layers, each of size 512, and ReLU activation. To evaluate image similarity we use LPIPS and as a metric for identity similarity, we use Arcface [13].

4.2. Model Evaluation

To evaluate our approach we sampled 1000 latent codes $\mathbf{w}$ from the generator G and measured the landmark loss

$$\mathcal{L}_{\text{L}}(\mathbf{w}) = ||(\mathcal{R} \circ \phi)(\mathbf{w}) - (\psi_L \circ G)(\mathbf{w})||^2. \tag{16}$$

Table 1. **Model evaluation.** Comparison of editing results in the latent spaces: $\mathcal{Z}$, $\mathcal{W}$, and $\mathcal{W}+$ of StyleGAN2 and 3. Performance is measured using different metrics, lower is better.

Model / latent space	$\mathcal{L}_L(\mathbf{w})$	$\mathcal{L}_L(\mathbf{w}_{\text{edit}})$	$\mathcal{L}_\phi$	$\mathcal{L}_\mathcal{R}$	$\mathcal{L}_{\text{ID}}$
sg2 / $\mathcal{Z}$	0.037	0.094	0.029	0.123	0.190
sg2 / $\mathcal{W}$	0.006	0.026	0.024	0.057	0.331
sg2 / $\mathcal{W}+$	0.008	0.036	0.058	0.181	0.019
sg3 / $\mathcal{Z}$	0.021	0.036	0.032	0.063	0.264
sg3 / $\mathcal{W}$	0.007	0.019	0.028	0.045	0.296
sg3 / $\mathcal{W}+$	0.009	0.021	0.071	0.160	0.034

We then perform a series of edits $\mathbf{w}_{\text{edit}} = \Omega_{\mathbf{w}}(\mathbf{q}_{\text{edit}})$ using the gradient-based method in (15) with Arcface for identity regularization with $\lambda_{\text{ID}} = 0.01$ For each edit, we measure the landmark loss $\mathcal{L}_{\text{L}}(\mathbf{w}_{\text{edit}})$ as well as three additional losses. First, we measure how well the edits results in the correct change in the prediction of the attribute vector with a metric $\mathcal{L}_\phi$ which we define as

$$\mathcal{L}_\phi = ||\phi(\mathbf{w}_{\text{edit}}) - \mathbf{q}_{\text{edit}})||^2. \tag{17}$$

Secondly, we measure how well the new ”ground truth” landmarks of the edited latent code agree with the target landmarks

$$\mathcal{L}_\mathcal{R} = ||\mathcal{R}(\mathbf{q}_{\text{edit}}) - (\psi_L \circ G)(\mathbf{w}_{\text{edit}})||^2. \tag{18}$$

Finally, we measure the identity loss $\mathcal{L}_{\text{ID}}$, between the original and edited images.

For this experiment, we used Dlib as the ”ground truth” landmark extractor ψ_L and evaluated the full 1024^2 resolution StyleGAN2 and 3 generators, both trained on the aligned FFHQ data set. We show the results in Table 1. The model was better at predicting landmarks in $\mathcal{W}$ and $\mathcal{W}+$ compared to $\mathcal{Z}$ space when measuring losses $\mathcal{L}_L(\mathbf{w})$ and $\mathcal{L}_L(\mathbf{w}_{\text{edit}})$.

We also observe that the identity loss $\mathcal{L}_{\text{ID}}$ is very low for $\mathcal{W}+$ space, however, $\mathcal{L}_\mathcal{R}$ is also dramatically higher, indicating that it is much harder to change the generated image in such a way that the extracted GT landmarks agree with the specified target when performing edits in $\mathcal{W}+$ space. The same point is supported by the $\mathcal{L}_\phi$ metric with is also substantially higher for $\mathcal{W}+$ space.

4.3. Identity Regularization

We performed a qualitative comparison between the linear (14) and gradient-based method (15), proposed in Section 3.3. Here we edited pose and smile using both methods and show the effect of adding identity regularization, *i.e.*, ArcFace, to the gradient-based method in Fig. 5. In the second column, it can be seen that the linear method is able to define directions in latent space which mostly change the target attribute, *i.e.*, pose or smile, however, we note

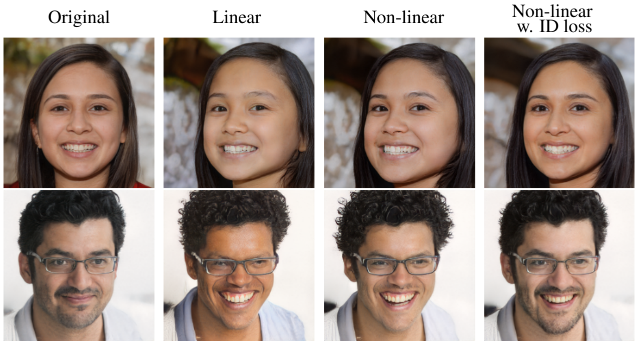


Figure 5. **The effect of identity regularization.** We observe that adding ArcFace to the loss function improves the identity preservation of two edits: rotation (top) and smile (bottom).

that the identity is not preserved well in the edit. This can be alleviated by the gradient-based method which defines a non-linear trajectory in latent space. Further, the gradient-based method in (15) allows for explicit identity regularization using ArcFace which substantially improves the degree of identity preservation for both pose and smile edits as can be seen in column 4 of Fig. 5.

4.4. Attribute Transfer

Our approach enables the transfer of attributes, such as pose or facial expression, from one image to another in a straightforward manner, while preserving other attributes such as identity and illumination. Given two latent codes, $\mathbf{w}_1$ and $\mathbf{w}_2$ with corresponding attribute vectors $\mathbf{q}_1$ and $\mathbf{q}_2$ we can transfer the pose and face shape from $\mathbf{w}_1$ to $\mathbf{w}_2$ by performing the edit $\widetilde{\mathbf{w}}_2 = \Omega_{\mathbf{w}_2}(\mathbf{q}_1)$. Here both $\mathbf{q}_1$ and $\mathbf{q}_2$ can be recovered using either the regressor ϕ or using the minimization procedure in (11). We demonstrate the results of our method in Fig. 6, where we changed the rotation and facial expression of three source images to match different target images, *i.e.*, transferring attributes from the target to the source, while preserving the identity in the source images.

4.5. Rotation and Translation with StyleGAN3

Our method is able to define trajectories in latent space corresponding to roll rotation as well as translations. As noted in [4] roll rotations and translations are a native part of the architecture of the StyleGAN3 generator and can be achieved by manipulating the Fourier features using the four parameters $(\sin\alpha, \cos\alpha, x, y)$ which are obtained from the first learned affine layer of the synthesis network. In comparison, our method can edit rotation and translation directly in the native $\mathcal{W}$ space of StyleGAN3. In Fig. 7, we qualitatively compare the effect of performing roll rotation and translation using our method to the effect of manipulating the Fourier features directly. We note translations look very similar with both methods. However, for roll rotations, we note that the axis of rotation is located in the middle of the left-hand image border when manipulating the Fourier features (see the location of the nose in Fig. 7), whereas, with our method, the axis of rotation is located at the center of the face.

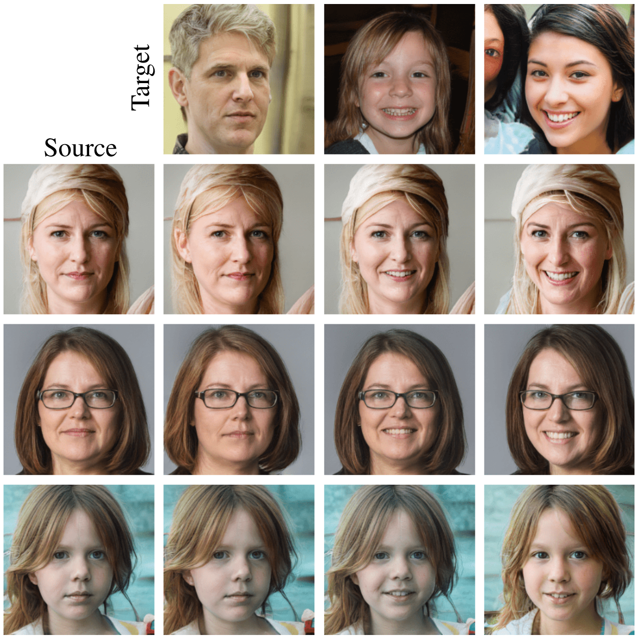


Figure 6. **Attribute transfer.** Our method can edit the rotation and expression of the source image (left column) to match the target image (top row) while preserving identity of the source.

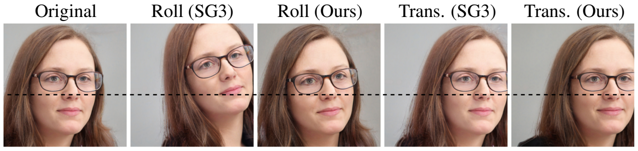
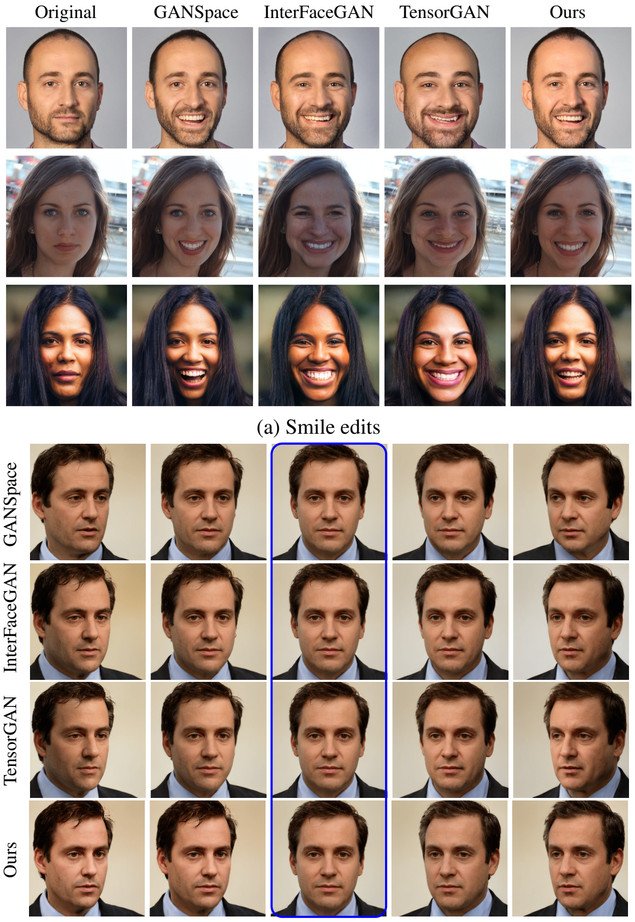


Figure 7. **Comparing our method to Fourier feature editing.** Our method finds a direction for roll rotation where the axis of rotation is at the center of the object. In comparison, manipulating the Fourier features results in an upward movement of the entire face since the axis of rotation is in the middle left border of the image. The vertical dotted line highlights the level of the nose for easier comparison.

4.6. Comparison with other Methods

We compared the editing directions corresponding to pose (yaw rotation) and smile with three off-the-shelf techniques for semantic editing: InterFaceGAN [38, 39], GANSpace [21], and TensorGAN [18, 19]. Although our method supports arbitrary 3D rotations in latent space, we focused on editing yaw rotations and smile since previous techniques have also been reported to enable these edits, enabling a direct comparison. A qualitative comparison of the edits to smile and yaw rotations generated by each of the



(a) Smile edits

(b) Yaw rotation edits applied to the original image shown in blue.

Figure 8. **Qualitative comparison to other methods.** We compare editing smile and yaw rotation using our method with equivalent edits using other off-the-shelf techniques.

four methods is shown in Fig. 8a and Fig. 8b respectively.

When evaluating the degree of identity preservation during the semantic edits it can be seen that our method is on par with the competing methods when performing yaw rotations and arguably better when editing smile.

4.7. Editing real images

Coupled with an encoder, our approach facilitates editing of real images. We qualitatively compared the projection and editing results when using our method in conjunction with e4e [45] and HyperStyle [5], respectively. The results are shown in Fig. 9. The two methods operate in different spaces, e4e project images into $\mathcal{W}+$ space while HyperStyle instead makes an initial prediction in $\mathcal{W}$ space and then fine-tunes the generator such that the prediction more faithfully reconstructs the target. Despite the fine-tuning of the generator it is not necessary to retrain the regressor when using HyperStyle for GAN Inversion.

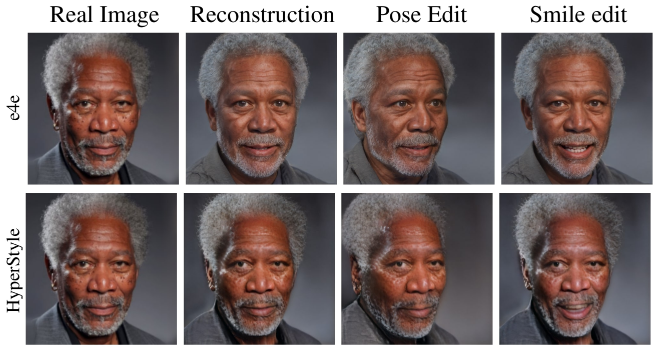


Figure 9. **Editing real images.** Qualitative comparison of projection and editing results when combining our method with two state-of-the-art encoders, e4e [45] and HyperStyle [5] respectively.

5. Conclusions

We presented a framework for highly interpretable image editing in pre-trained 2D GANs. Our framework provides an efficient method to find trajectories in the latent space of GANs which change the generated images according to camera, orientation, and shape parameters. This enables the discovery of trajectories in the latent space corresponding to arbitrary transformations of shape and orientation of the generated images.

In summary, we first used NRSfM to derive a sparse 3D model on the domain of the generator. We then trained a regressor to relate the 3D model to the latent space. We then proposed two methods for using the regressor for semantic editing: a linear method, and a gradient-based method. The latter is similar to the iterative editing algorithm in [46], however, we integrate explicit identity regularization which improves identity preservation.

Our method provides an efficient framework for manipulating the 3D structure of objects generated by 2D GANs. Compared to other methods, our approach is fast compared to existing frameworks for training explicitly 3D aware GANs [17, 31, 48] and compared to [43] our method is lightweight and able to perform rotations and edits to face shape without the need for a 3D morphable model. Since our method only requires access to a landmark extractor trained on the same domain as the generator, our approach does not require any additional training data and can be trained in a fully self-supervised fashion. Further, our approach does not require retraining of the generator or any changes to the generator architecture.

As to limitations, our method allows for adjustments to the position, orientation as well as non-rigid deformation of the face shape of the generated images. Since our method only captures the 3D orientation and face shape our method is not able to add or remove face accessories, eye-glasses, earrings, and hats nor change the skin tone or hair color. Overcoming those limitations is an avenue for future work.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 8293–8302, Seattle, WA, USA, Jun 2020. IEEE. 1, 3

[2] Rameen Abdal, Peihao Zhu, Niloy J. Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Trans. Graph.*, 40(3), May 2021. 1, 3

[3] Yuval Alaluf, Or Patashnik, and Daniel Cohen-Or. Restyle: A residual-based stylegan encoder via iterative refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021. 3

[4] Yuval Alaluf, Or Patashnik, Zongze Wu, Asif Zamir, Eli Shechtman, Dani Lischinski, and Daniel Cohen-Or. Third time's the charm? image and video editing with stylegan3. *Advances in Image Manipulation Workshop - ECCV 2022*, Jan 2022. 3, 7

[5] Yuval Alaluf, Omer Tov, Ron Mokady, Rinon Gal, and Amit H. Bermano. Hyperstyle: Stylegan inversion with hypernetworks for real image editing. In *Proc. CVPR*, 2022. 1, 3, 8

[6] Peter Baylies. Stylegan encoder - converts real images to latent space. https://github.com/pbaylies/stylegan-encoder/, 2019. 3

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proc. SIGGRAPH*, pages 187–194, 1999. 3

[8] Sami S. Brandt and Hanno Ackermann. Non-rigid structure-from-motion by rank-one basis shapes, Apr 2019. arXiv: 1904.13271. 2, 4, 5

[9] Sami Sebastian Brandt, Hanno Ackermann, and Stella Grasshof. Uncalibrated non-rigid factorisation by independent subspace analysis. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 569–578, 2019. 4, 5

[10] C. Bregler, A. Hertzmann, and H. Biermann. Recovering non-rigid 3d shape from image streams. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No.PR00662)*, volume 2, page 690–696. IEEE Comput. Soc, 2000. 3

[11] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1

[12] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[13] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proc. CVPR*, pages 4690–4699, 2019. 5, 6

[14] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. *CoRR*, abs/1508.06576, 2015. 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, page 2672–2680. Curran Associates, Inc., 2014. 1

[16] Stella Graßhof and Sami Sebastian Brandt. Tensor-based non-rigid structure from motion. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, page 2254–2263. IEEE, Jan 2022. 2, 4, 5

[17] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 3, 8

[18] René Haas, Stella Graßhof, and Sami Sebastian Brandt. Tensor-based subspace factorization for stylegan. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–8, Los Alamitos, CA, USA, dec 2021. IEEE Computer Society. 7

[19] René Haas, Stella Graßhof, and Sami Sebastian Brandt. Tensor-based emotion editing in the stylegan latent space. *arXiv:2205.06102 [cs]*, May 2022. Accepted for poster presentation at AI4CC @ CVPRW. 7

[20] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *Proc. ICCV*, Jul 2017. 2

[21] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. In *Proc. NeurIPS*, 2020. 1, 3, 7

[22] Sebastian Hoppe Nesgaard Jensen, Mads Emil Brix Doest, Henrik Aanæs, and Alessio Del Bue. A benchmark and evaluation of non-rigid structure from motion. *International Journal of Computer Vision*, 129(4):882–899, Apr 2021. 3

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *Proc. ICLR*, Feb 2018. 1

[24] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. In *Proc. NeurIPS*, 2020. 1, 2

[25] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. In *Proc. NeurIPS*, 2021. 1, 2, 3, 6

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. CVPR*, pages 4396–4405, 2019. 1, 2, 6

[27] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 1, 2, 3, 6

[28] Davis E. King. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.*, 10:1755–1758, Dec. 2009. 6

[29] Chen Kong and Simon Lucey. Deep Non-Rigid Structure From Motion. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1558–1567, Oct. 2019. ISSN: 2380-7504. 3

[30] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, and et al. Mediapipe: A framework for building perception pipelines. *arXiv:1906.08172 [cs]*, Jun 2019. arXiv: 1906.08172. 6

[31] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 11448–11459, Nashville, TN, USA, Jun 2021. IEEE. 3, 8

[32] Yotam Nitzan, Rinon Gal, Ofir Brenner, and Daniel Cohen-Or. Large: Latent-based regression through gan semantics. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19217–19227, 2022. 1

[33] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021. 1

[34] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible conditional gans for image editing. In *NIPS 2016 Workshop on Adversarial Training*, Nov 2016. 3

[35] Yipeng Qin Rameen Abdal and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proc. ICCV*, pages 4431–4440, 2019. 2, 3

[36] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 1, 3

[37] Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021. 3

[38] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *CVPR*, 2020. 2, 7

[39] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *TPAMI*, 2020. 1, 2, 7

[40] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *CVPR*, 2021. 3

[41] Vikramjit Sidhu, Edgar Tretschk, Vladislav Golyanik, Antonio Agudo, and Christian Theobalt. Neural Dense Non-Rigid Structure from Motion with Latent Space Constraints. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[42] Nurit Spingarn, Ron Banner, and Tomer Michaeli. GAN Steerability without optimization. In *International Conference on Learning Representations*, 2021. 3

[43] Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. Stylerig: Rigging stylegan for 3d control over portrait images, cvpr 2020. In *Proc. CVPR)*. IEEE, june 2020. 1, 3, 8

[44] Carlo Tomasi and Takeo Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision*, 9(2):137–154, 1992. 3

[45] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics*, 40(4):133:1–133:14, July 2021. 1, 2, 3, 8

[46] Hui-Po Wang, Ning Yu, and Mario Fritz. Hijack-gan: Unintended-use of pretrained, black-box gans. In *Proc. CVPR)*, page 10, 2021. 2, 5, 8

[47] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proc. CVPR*, Dec 2020. 1, 3

[48] Doğa Yılmaz, Furkan Kınlı, Barış Özcan, and Furkan Kıraç. [re] lifting 2d styleGAN for 3d-aware face generation. In *ML Reproducibility Challenge 2021 (Fall Edition)*, 2022. 3, 8

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc CVPR*, 2018. 5

[50] Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. In-domain gan inversion for real image editing. In *Proc. ECCV*, 2020. 2, 3