GlassesGAN: Eyewear Personalization using Synthetic Appearance Discovery and Targeted Subspace Modeling[*]

Richard Plesh[1] Peter Peer[2] Vitomir Struc[2]
[1] Clarkson University, USA [2] University of Ljubljana, Slovenia
https://github.com/pleshro/GlassesGAN_release

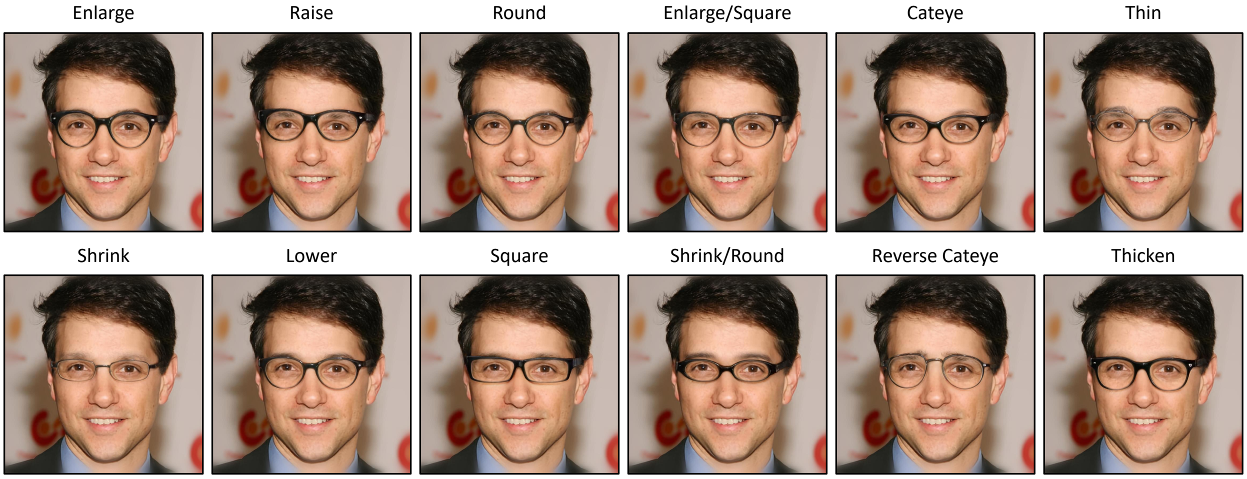

Figure 1. This paper introduces GlassesGAN, an innovative approach to image editing capable of generating continuously tunable, multi-attribute, and photo-realistic editing of eyeglasses by leveraging a novel method for modeling sub-spaces in the StyleGAN2 latent space. The presented (1024 × 1024) examples show editing results for twelve different tuning attributes. Best viewed zoomed-in.

Abstract

We present GlassesGAN, a novel image editing framework for custom design of glasses, that sets a new standard in terms of output-image quality, edit realism, and continuous multi-style edit capability. To facilitate the editing process with GlassesGAN, we propose a Targeted Subspace Modelling (TSM) procedure that, based on a novel mechanism for (synthetic) appearance discovery in the latent space of a pre-trained GAN generator, constructs an eyeglasses-specific (latent) subspace that the editing framework can utilize. Additionally, we also introduce an appearance-constrained subspace initialization (SI) technique that centers the latent representation of the given input image in the well-defined part of the constructed subspace to improve the reliability of the learned edits. We test GlassesGAN on two (diverse) high-resolution datasets (CelebA-HQ and SiblingsDB-HQf) and compare it to three state-of-the-art baselines, i.e., InterfaceGAN, GANSpace, and MaskGAN. The reported results show that GlassesGAN convincingly outperforms all competing techniques, while offering functionality (e.g., fine-grained multi-style editing) not available with any of the competitors. The source code for GlassesGAN is made publicly available.

1. Introduction

Consumers are increasingly choosing the convenience of online shopping over traditional brick-and-mortar stores [3]. For the apparel industry, which traditionally relied on individuals being able to try on items to suit their taste and body shape before purchasing, the shift to digital commerce has created an unsustainable cycle of purchasing, shipping, and returns. Now, an estimated 85% of manufactured fashion items end up in landfills each year, largely due to consumer returns and unsatisfied online customers [35].

In response to these challenges, the computer vision community has become increasingly interested in virtual-try-on (VTON) techniques [3, 6, 9, 11] that allow for the development of virtual fitting rooms, where consumers can try on clothing in a virtual setting. Furthermore, such techniques also give users the flexibility to explore custom designs and personalize fashion items by rendering them photo-realistically in the provided input image.

While considerable progress has been made recently in

[*] Supported by ARRS J2-2501(A), the Fulbright Scholarship Fund, the Center for Identification Technology Research, and the National Science Foundation under Grant No. 1650503.

image-based virtual try-on techniques for clothing and apparel that do not require (costly) dedicated hardware and difficult-to-acquire 3D annotated data [5,6,11,18,48], most deployed solutions for virtual **eyewear try-on** still largely rely on traditional computer graphics pipelines and 3D modeling [1,29,30,49,51]. Such 3D solutions provide convincing results, but save for a few exceptions, e.g., [14], are only able to handle predefined glasses and do not support custom designs and eyewear personalization. Although work has also been done with 2D (image) data only, relevant research for virtual eyewear try-on has mostly focused on editing technology (facilitated by Generative Adversarial Network - GANs [10]) capable of inserting glasses into an image [15, 27, 34, 37, 44]. The images these methods generate are often impressive, but adding eyewear with finely tunable appearance control still remains challenging.

In this work, we address this gap by introducing GlassesGAN, a flexible image editing framework that allows users to add glasses to a diverse range of input face images (at a high-resolution) and control their appearance. Distinct from existing virtual try-on work in the vision literature, the goal of GlassesGAN is not to try on existing glasses, but rather to allow users to explore custom eyewear designs. GlassesGAN is designed as a GAN inversion method [45], that uses a novel *Targeted Subspace Modeling (TSM)* technique to identify relevant directions within the latent space of a pre-trained GAN model that can be utilized to manipulate the appearance of eyeglasses in the edited images. A key component of GlassesGAN is a new *Synthetic Appearance Discovery (SAD)* mechanism that samples the GAN latent space for eyeglasses appearances, without requiring real-world facial images with eyewear. Additionally, we propose an *appearance-constrained subspace initialization procedure* for the (inference-time) editing stage, which helps to produce consistent editing results across a diverse range of input images. We evaluate GlassesGAN in comprehensive experiments over two test datasets and in comparison to state-of-the-art solutions from the literature, with highly encouraging results.

In summary, our main contributions in this paper are:

- We present GlassesGAN, an image editing framework for custom design of eyeglasses in a virtual try-on setting that sets a new standard in terms of output image quality, edit realism, and continuous multi-style edit capability, as illustrated in Figure 1.
- We introduce a Synthetic Appearance Discovery (SAD) mechanism and a Targeted Subspace Modeling (TSM) procedure, capable of capturing eyeglasses-appearance variations in the latent space of GAN models using glasses-free facial images only.
- We introduce a novel initialization procedure for the editing process that improves the reliability of the facial manipulations across different input images.

2. Related work

In this section, we briefly review existing work needed to provide context for GlassesGAN. The reader is referred to some of the excellent surveys on generative models [43], image editing [38,45] and virtual try-on [3,9,19] for a more comprehensive coverage of relevant areas.

Generative Adversarial Networks (GANs) represent a class of generative models capable of synthesizing realistic, high-quality imagery [10] and consist of generative and discriminative sub-networks learned with competing objectives [10]. Recent advances in GAN design and associated training procedures have led to considerable progress in various areas, including image-to-image translation [16, 26, 33, 42, 53], image attribute manipulation [15,27,37,41,46] as well as virtual try-on and fashion-related applications [6, 8, 11, 17]. Modern GAN models, such as StyleGAN (v1–v3) [22–24], have had particular success in generating realistic high-resolution (facial) images and facilitate corresponding editing solutions.

Latent Space Image Editing techniques alter attributes in the given input image by encoding the image in the GAN latent space, modifying the embedding, and then decoding the modified embedding [25, 28, 31, 32, 37, 47]. While these types of methods can be very flexible, they typically suffer from a trade-off between editability, image consistency, distortion, and perceptual quality [39]. Additionally, the entanglement between different attributes in the generated images limits the locality of edits [36, 37]. To mitigate such shortcomings, some researchers bypass the trade-off by blending the original image with the edited output image at strategic locations [32], something we also follow with the proposed GlassesGAN framework in this work.

Glasses VTON. Recent works have had success creating VTON systems that rely upon detailed 3D modeling of the eyeglasses and/or the head [1, 7, 29, 30, 49–51]. While some implementations have impressive edit realism, every additional eyeglass style (and person) requires a new 3D model. As a consequence, these techniques scale poorly to new eyeglasses, rarely contain the capability to make (continuous) edits to the eyeglasses, and sometimes require an initial 3D scan of the face/head to be applicable.

To address such shortcomings, many recent methods try to avoid 3D data altogether and exploit advances in face image editing. These methods include latent space editing solutions, [15,25,36], but also other editing strategies, [12,27], capable of adding glasses to an input face image. While latent space editing techniques can be performed on 2D facial images and provide a realistic edit to the images, they have substantial problems with preserving identity throughout the edit, isolating the edit to the eyeglasses, and, prior to GlassesGAN, offered no multi-style personalization.

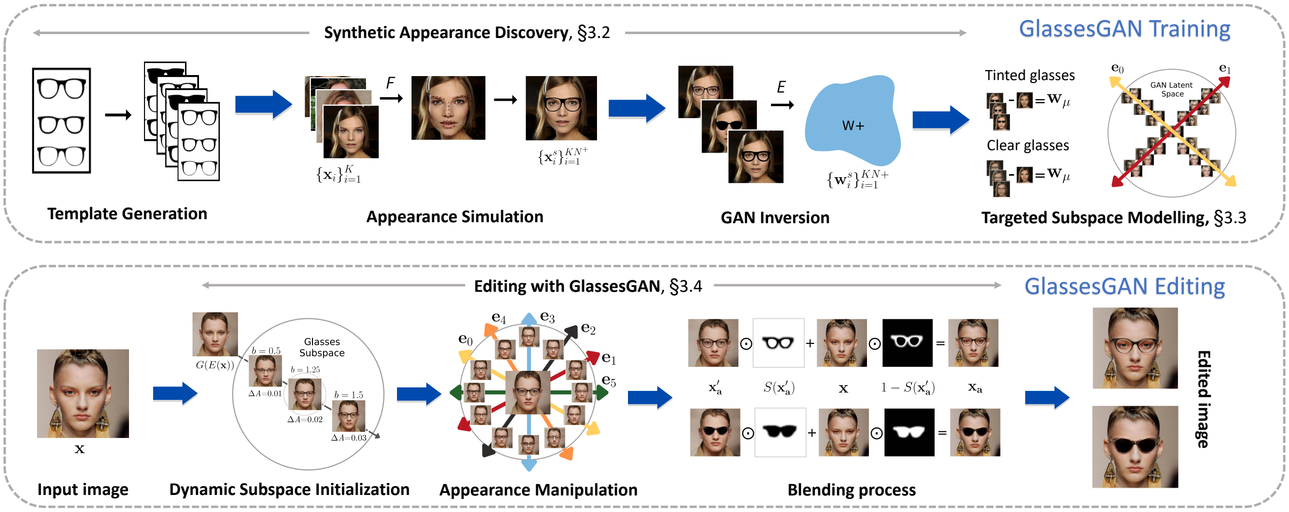

Figure 2. **Overview of the GlassesGAN framework.** GlassesGAN learns continuous multi-style edits through a novel GAN latent space sampling technique (synthetic appearance discovery) that first embeds augmented images into the latent space of a pretrained GAN generator and then captures the data distribution using the Karhunen-Loève Transform. During editing, the framework dynamically initializes the latent vector in the center of the glasses subspace for greater edit consistency and then modifies different glasses attributes as desired.

3. Methodology

In this section, we present the main contribution of this work: GlassesGAN, a novel image editing framework that allows for the personalization of eyeglasses in a virtual setting, i.e., with visual feedback to the user.

3.1. Overview of GlassesGAN

Problem formulation. Given an input face image $\mathbf{x} \in \mathbb{R}^{m \times n \times 3}$ and some desired semantics a (i.e., appearance of glasses), the goal of GlassesGAN is to construct a mapping $\psi_a : \mathbf{x} \mapsto \mathbf{x}_a \in \mathbb{R}^{m \times n \times 3}$, such that the edited output image $\mathbf{x}_a$ incorporates the semantics a in a realistic and visually convincing manner, while preserving the original image content as much as possible, e.g., facial appearance, background, and identity. A few illustrative examples of such edited images $\mathbf{x}_a$ are presented in Figure 1.

Many state-of-the-art image-editing techniques implement the mapping ψ_a through so-called GAN inversion approaches [37,45], where the input image $\mathbf{x}$ is first embedded into the latent space of a pretrained GAN generator G, thus, resulting in a latent representation w. This latent code is then modified, i.e., $\psi_a^{latent} : w \mapsto w_a$, such that the generated image $\mathbf{x}_a = G(w_a)$ adheres as closely as possible to the facial editing constraints. GlassesGAN follows this general latent-space editing framework, but in contrast to prior work: (i) does not require a dataset with the attribute a present to define ψ_a^{latent}, and (ii) learns latent space manipulations that enable *continuous multi-style* changes to a.

GlassesGAN design. A high-level overview of GlassesGAN in presented in Fig. 2. Central to the editing ability of the framework are two novel components, i.e., (i) a mechanism for *Synthetic Appearance Discovery* (SAD) that allows us to sample target appearances of faces with various styles of glasses ($\mathbf{x}_a$) and their corresponding GAN latent codes **without actual real-world data** (§3.2), and (ii) a *Targeted Subspace Modeling* (TSM) approach (§3.3) that based on the sampled representations, determines the latent editing directions via the Karhunen-Loève Transform [20].

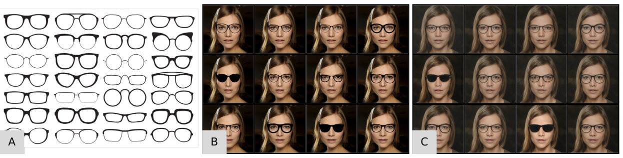

Figure 3. **Illustration of the SAD steps.** From left to right: (a) the initial (binary) glasses templates, (b) faces with superimposed templates, (c) re-renderings after latent space embedding.

The identified latent directions correspond to different types of eyeglasses edits and can be applied to an input image's latent code as desired. To avoid problems with the latent space manipulations, we also propose a novel dynamic *Subspace Initialization* (SI) procedure (§3.4) that ensures that the generated edits are semantically meaningful. To produce the final output $\mathbf{x}_a$, we finally use a blending operation with the original image $\mathbf{x}$, which helps to preserve identity and to improve the locality of the edits.

3.2. Synthetic Appearance Discovery

The majority of existing latent-space editing techniques require (paired or unpaired) data with and without the desired semantics a to be able to learn the mapping ψ_a, e.g., [37, 47]. Since our goal is to provide fine-grained control over the appearance of glasses and suitable datasets for this purpose are not publicly available, we propose a Synthetic Appearance Discovery (SAD) mechanism to mitigate this problem. Details on the mechanism are given below.

Step 1: Template generation. We work under the assumption that only facial images $\mathbf{x}$ without glasses are available. To generate paired data with and without glasses, we simulate the presence of eyewear by superimposing hand-drawn binary masks $\mathbf{b}$ (*glasses templates* hereafter) over the input images. We start this process with a collection of N initial masks (see Figure 3(a)), which we augment using morphological modifications, such as dilation and erosion, to expand the variability in the set of glasses templates.

Step 2: Appearance simulation. Next, we add the augmented set of N^+ binary masks to each input image $\mathbf{x}$, resulting in facial images with an artificial cut-and-paste look $\mathbf{x}^s$, as illustrated in Figure 3(b). The addition of the glasses is implemented based on a facial landmarking procedure F that allows us to place the glasses templates on the faces in such a way that the temples of the head overlap with the outer points of the glasses frames (see also Figure 2). Since N^+ glasses templates are available, this step results in a set of N^+ images $\{\mathbf{x}_i^s\}_{i=1}^{N^+}$ for each given input image.

Step 3: GAN inversion. Finally, we embed the augmented images $\{\mathbf{x}_i^s\}_{i=1}^{N^+}$ in the latent space of the generator G to obtain the corresponding latent codes $\{\mathbf{w}_i^s\}_{i=1}^{N^+}$. A pretrained StyleGAN2 model is used as the generator for GlassesGAN with the extended (512×18 dimensional) W^+ latent space [24]. We use an encoder-based approach for the GAN inversion, where the latent codes are computed as $\mathbf{w} = E(\mathbf{x})$ and E represents the encoding operation. This last encoding step relies on the properties of pre-trained generator models, which are known to interpret image artifacts and binary occlusions in a semantically meaningful manner. As a result, the computed binary codes, simulate glasses with realistic appearance, and even add shadowing and specular reflections when re-rendered through the generator, i.e., $\mathbf{x}_s' = G(\mathbf{w}^s)$, as shown in Figure 3(c).

If we assume a training set of K glasses-free facial images, the SAD mechanism results in a dataset of KN^+ latent codes that capture the variability induced by the presence of eyeglasses and are used in the targeted subspace modeling (TSM) procedure, described in the next section.

3.3. Targeted Subspace Modeling

To facilitate continuous multi-style editing in the latent space, we introduce a Targeted Subspace Modeling (TSM) procedure, capable of identifying relevant latent space directions that, when traversed, result in visually meaningful modifications in the appearance of eyeglasses. Assume that: (i) a training set of K facial images without glasses is available, (ii) that N^+ latent codes $\{\mathbf{w}_i^s\}_{i=1}^{N^+}$ have been computed with the SAD for each training image, and (iii) that the center $\mathbf{w}_\mu^s = (1/N^+)\sum_{i=1}^{N^+}\mathbf{w}_i^s$ of these latent codes has been determined. TSM then first computes a differential latent code for each of the K images, i.e.:

$$\mathbf{\Delta W} = [vec(\mathbf{w}_1^s - \mathbf{w}_\mu^s), \dots, vec(\mathbf{w}_{N^+}^s - \mathbf{w}_\mu^s)], \quad (1)$$

where d is the dimensionality of the W^+ latent space (i.e., $d = 512\cdot18$) and $vec(\cdot)$ denotes a vectorization operator, and then aggregates the differentials over the training data:

$$\mathbf{W} = [\mathbf{\Delta W}_1, \mathbf{\Delta W}_2, \dots, \mathbf{\Delta W}_K] \in \mathbb{R}^{d\times KN^+}. \quad (2)$$

The latent code differences in $\mathbf{W}$ capture the appearance variations of glasses, introduced to the training images by the SAD mechanism, and span a *glasses subspace* within the latent space of the generator. As we show in the experimental section, the differential formulation introduced above also allows us to model variations of different types of glasses (e.g., with clear and tinted lenses) using a single latent subspace. This subspace is identified by solving the eigenproblem given by the Karhunen-Loève Transform:

$$\mathbf{\Sigma}\mathbf{e}_i = \lambda_i\mathbf{e}_i, \;\; i = 1, 2, \dots, d', \quad (3)$$

where $\mathbf{\Sigma} = \mathbf{W}\mathbf{W}^T$ is an image-conditioned intra-class scatter matrix, and $d' \leq d$. The leading eigenvectors corresponding to non-zero eigenvalues, i.e., $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{d'}] \in \mathbb{R}^{d\times d'}$, define the (orthonormal) principal axes of the glasses subspace in W^+ and represent the basis for the image editing procedure of GlassesGAN.

As part of TSM, we also compute a difference vector $\mathbf{w}_\mu$ between the latent code $\mathbf{w}$ of each glasses-free training image and the centroid (i.e, mean vector) of the latent codes corresponding to the KN^+ glasses-augmented samples $\{\mathbf{w}_i^s\}_{i=1}^{KN^+}$. This difference vector is required for the initialization of the editing procedure.

3.4. Editing with GlassesGAN

The editing procedure implemented for GlassesGAN consists of three main parts, as detailed below.

Part 1: Latent Code Editing. Given a glasses-free input image $\mathbf{x}$ and its corresponding latent code $\mathbf{w} = E(\mathbf{x})$, computed with a pre-trained encoder E, we alter the initial latent code $\mathbf{w}$ by traversing the principal subspace axes in $\mathbf{E}$ using the following expression:

$$\mathbf{w}_a' = vec(\mathbf{w}) + b\cdot vec(\mathbf{w}_\mu) + m\cdot\mathbf{e}_i, \quad (4)$$

where $i \in \{1, 2, \dots, d'\}$, $m \in [-\infty, \infty]$ is a real-valued scalar that controls the strength of the edits (*editing magnitude* hereafter), b is a weighting parameter that is set dynamically as part of the initialization procedure (described in Part 2), and the final 512×18 latent representation $\mathbf{w}_a$ of the initial edited output image $\mathbf{x}_a' = G(\mathbf{w}_a)$ is computed as $\mathbf{w}_a = vec^{-1}(\mathbf{w}_a')$. Each principal axis $\{\mathbf{e}_i\}_{i=1}^{d'}$

controls a specific attribute (or style) of the glasses, while tuning the magnitude m allows for continuous appearance changes w.r.t said attribute. The addition of the average difference code $\mathbf{w}_\mu$ serves as an initialization step that moves $\mathbf{w}$ into the well-defined part of the computed subspace, as shown in Figure 2. If $\mathbf{w}_\mu$ is computed based only on latent codes corresponding to specific styles of glasses (e.g., clear or tinted), then this code can also be used to define the initial appearance of the eyewear added to the image.

Part 2: Dynamic Subspace Initialization. It is important to note that the magnitude of the weighting parameter b in Eq. (4), has significant downstream effects on later style edits, with improper values leading to eyeglasses that are poorly rendered or even non-existent. Similarly to prior work [32, 37]) we observed that the use of a fixed value of b produces highly inconsistent edits across different samples. We hypothesize that this is because some samples are farther from the relevant part of the latent space than others. If the value of b is too small for a particular sample, the embedding never enters the glasses subspace and, as a result, the glasses never (properly) appear. To address this issue, we propose a *Subspace Initialization* (SI) procedure that dynamically adjusts the value of b on a per-sample basis and ensures consistent editing results when using fixed, predefined style editing magnitudes m. Central to the initialization operation is the realization that the modified latent code w_a is near the center of the glasses subspace when the frames of the glasses in the corresponds image $G(w_a)$ cover a certain fraction ΔA of the overall image area. The initialization process therefore sets m to zero, iteratively samples a range of values of b from 0.5 to 1.5, generates an output image, subjects it to a face parser S capable of segmenting the face from the glasses, and finally selects the optimal value of b, such that the frames in $G(w_a)$ cover a relative area as close to ΔA as possible, as shown in Figure 2.

Part 3: Blending. As illustrated in Figure 2, in the last step, we finally blend the glasses region of the edited image $\mathbf{x}'_a$ with the original image $\mathbf{x}$ to improve the preservation of identity and, thus, compute the final output $\mathbf{x}_a$. The blending mask comes from the face parser S applied to x'_a. The edges of the mask are tapered using Gaussian blur to smooth the boundary between the original and edited images. In the case of clear glasses two separate Gaussian blur operations for the interior and exterior of the glasses frames are used to better preserve the eyes of the original image.

4. Experiments And Results

In this section, we now present the experiments conducted to highlight the characteristics of GlassesGAN.

4.1. Datasets and Experimental Splits

Three face datasets with diverse characteristics are used in the experiments with GlassesGAN, as summarized in Table 1, i.e.: FFHQ [24], CelebA-HQ [21], and SiblingsDB-HQf [40]. The datasets represent standard datasets used when evaluating image editing techniques and were, therefore, also selected for the experiments in this work [24, 32]:

Table 1. **Summary of the experimental datasets and data splits.**

Dataset	**Resolution**	**Purpose**$^\diamond$	**#Train. Img.**†	**#Test Img.**	**Variability**‡
FFHQ [24]	1024×1024	TR (G,E)	$70,000$	n/a	A, ET, G, B
CelebA–HQ [21]	1024×1024	TR (S), Q, TS	1000	1000	A, ET, B, G, AC
SiblingsDB-HQf [40]	4256×2832	Q, TS	n/a	163	A, G

$^\diamond$ TR – training, Q – qualitative evaluation, TS – quantitative evaluation (testing), n/a – not applicable.

† The number of training images reported includes both training and validation data.

‡ A – age, ET – ethnicity, G – gender, B – background, AC – accessories.

- **FFHQ** [24] consists of $70,000$ facial images of 1024×1024 pixels in size and was acquired from Flickr. Due to the unconstrained nature of the collection procedure, the datasets exhibits variability across various factors. FFHQ is used to train the generator G and image encoder E in our experiments.
- **CelebA-HQ** [21] contains high-quality facial images at a resolution of 1024×1024 pixels with considerable appearance variability. 1000 sampled images are used for training (with SAD and TSM) in our experiments, and a non-overlapping subset of 1000 diverse test images is used for the quantitative evaluation.
- **SiblingsDB-HQf** [40] contains 184 frontal facial images of 92 sibling pairs captured at a resolution of 4256×2832. The dataset was acquired in front of a homogenous background and under diffuse illumination. This dataset is used exclusively for testing to demonstrate the generalization capabilities of GlassesGAN across datasets. After removing duplicates and excluding problematic samples, 163 image are left for the quantitative part of the evaluation.

We note that the training and test data is kept disjoint in all experiments, both in terms of images and subjects identities.

4.2. Implementation Details and Runtime

For the implementation of GlassesGAN, we use StyleGAN2 at resolution 1024×1024 trained on images from FFHQ [24] as the generator G of our framework and the e4e [39] encoder E again trained on FFHQ for inverting images into StyleGAN's latent space. For face detection and identifying facial landmarks we utilize the 68-point landmark model provided in the dlib package. We adopt the DatasetGAN [52] framework with 7 manual annotated data samples to generate synthetic training data for the face parser and then learn a DeeplabV2 model to serve as the parser S in our experiments [2]. We construct a $d' = 6$ dimensional subspace from the CelebA-HQ training images, and use $N = 28$ glasses templates for TSM. The image area threshold ΔA is set to 0.02 based on preliminary experiments. With the current implementation using an RTX 3090 GPU, adding glasses and applying an edit to an image

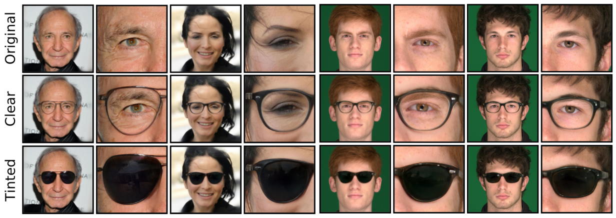

Figure 4. **Addition of initial glasses of a certain style.** GlassesGAN is able to render and edit glasses in different styles. The bottom two rows shows examples of the initialization with (average/initial) clear (middle) and tinted (bottom) glasses added to the (CelebA-HQ and SiblingsDB-HQf) input images on the top.

requires 4.8s on average (estimated over 100 test images). The addition of the subspace initialization to the pipeline costs an additional 12.7s. However, an efficient parallel implementation of GlassesGAN is expected to allow for real-time editing capability. Additional implementation details can be found in the publicly released source code.

4.3. Qualitative Results

To demonstrate the capabilities of GlassesGAN, we first present a series of visual results that illustrate: (i) the addition of two different types of initial eyeglasses to a face image, (ii) the tuning of eyeglasses appearance with respect to different attributes, (iii) sequentially chaining of eyeglass style edits, and (iv) edits to eyeglass frame color.

Adding initial glasses. The initialization procedure of GlassesGAN requires that a starting point is chosen in the glasses subspace via $\mathbf{w}_\mu$ in Eq. (4). This starting point defines the initial appearance and shape of the rendered glasses and can be varied to achieve different results, i.e., different initial styles of glasses. In Figure 4 we show a number of qualitative examples, where initial (average) clear and tinted glasses were added to the input images. As can be seen, GlassesGAN is able to add glasses to input images with diverse appearances (i.e., varying gender, age, background, color characteristics, etc.) and automatically consider facial alignment, shadowing, the boundary with the hair, and reflections in the frames and lenses. Additionally, we see that the blending procedure helps to maintain the fine image details while still preserving identity.

Editing different attributes. Using the TSM procedure, GlassesGAN identifies a number of latent subspace directions that can be traversed to alter the appearance of the generated glasses. In Figure 5 we present a few visual examples where the initial glasses in the middle column are altered (continuously) in six different directions. Each of the rows corresponds to changes along one subspace axis from Eq. (3). Because the TSM procedure is unsupervised, we subjectively assign human-interpretable attributes to these directions, which impact the following aspects of the added glasses: size, height/position, squareness, roundness, cateye appearance, and thickness. Note that each of the edits is visually convincing and creates distinct appearances.

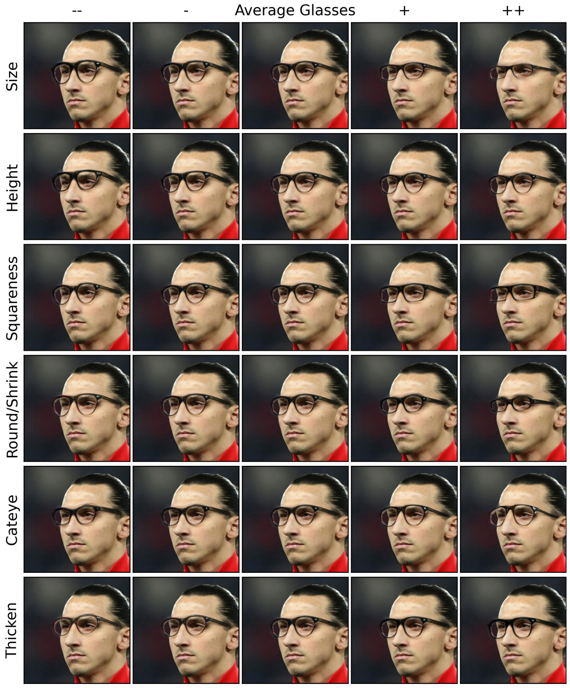

Figure 5. **Continuous multi-style edits.** Each row shows a separate edit on a challenging pose that starts from the (initialized) image in the middle and modifies one aspect of the glasses in a given direction. Results corresponding to the first six subspace axes (top to bottom) identified through the TSM procedure are presented.

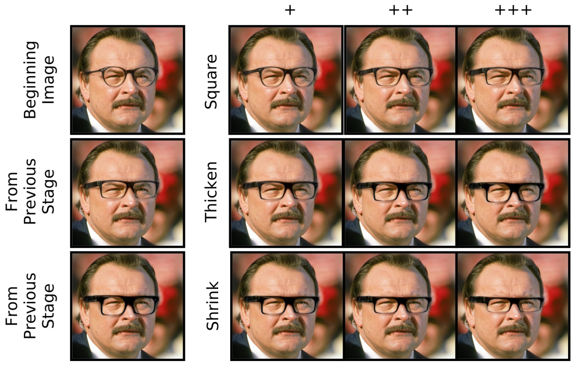

Figure 6. **Example of sequentially chained edits**. GlassesGAN allows to chain edits in different latent subspace directions without affecting the realism of the results or introducing artifacts.

Multiple chained edits. Next, we show that the latent subspace directions exploited by GlassesGAN are disentangled enough (due to the orthogonality of the learned subspace) to allow multiple chained edits to an image. For example, in Figure 6, we show that eyeglasses can be sequentially squared, thickened, and then shrunk. This chained editing procedure allows for the generation of unique appearances of glasses and fine-grained control over the editing procedure - a characteristic unique to our framework.

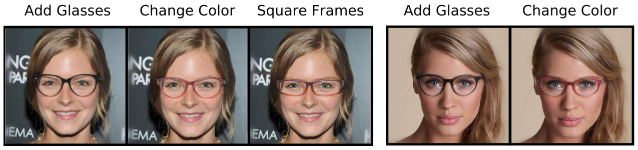

Figure 7. **Examples of color-related edits with GlassesGAN.** The example of the left adds glasses to the face, changes the color, and then squares frames using the learned edit vectors. The right example first adds the glasses and then changes the color.

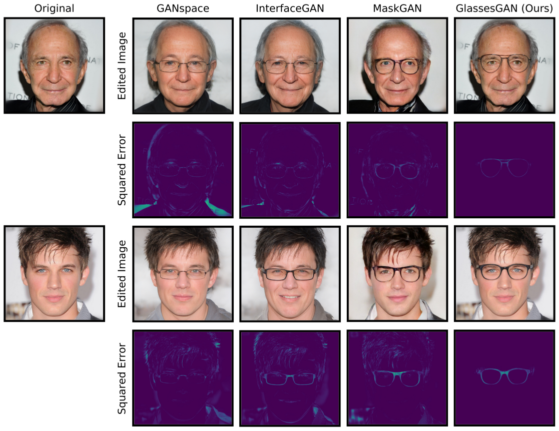

Figure 8. **Comparison to the state-of-the-art.** The examples show that GlassesGAN generates convincing results with minimal (or no) changes in identity. Squared pixel differences between the originals and edits are shown to highlight modified image areas.

Color change. In Figure 7 we demonstrate the ability of GlassesGAN to also capture attributes beyond the shape of the glasses. Specifically, by using colored augmentations for the glasses templates used in the SAD mechanism, we obtain frame-color control that is (reasonably well) disentangled from our suite of frame shape edits. This speaks of the flexibility of the framework and points to the potential for supporting further editing attributes if required.

Table 2. **Comparison to the state-of-the-art.** GlassesGAN outperforms all baselines on both test datasets across nearly all performance indicators by a wide margin. The arrow (↓↑) indicates if lower or higher scores imply better performance.

Method	**CelebA-HQ**		
	MSE (↓)	IDS (↓)	FID (↓)
InterfaceGAN [37]	0.0173 ± 0.0058	0.5789 ± 0.1026	58.15
MaskGAN† [27]	0.0149 ± 0.0064	0.6568 ± 0.0975	53.11
GANSpace [15]	0.0153 ± 0.0064	0.4842 ± 0.1060	40.45
GlassesGAN (ours)	$\mathbf{0.0029 \pm 0.0009}$	$\mathbf{0.1707 \pm 0.0625}$	**26.02**
Method	**SiblingsDB-HQF**		
	MSE (↓)	IDS (↓)	FID (↓)
InterfaceGAN [37]	0.0099 ± 0.0022	0.5780 ± 0.0790	80.64
GANSpace [15]	0.0085 ± 0.0030	0.5047 ± 0.0883	60.79
GlassesGAN (ours)	$\mathbf{0.0029 \pm 0.0007}$	$\mathbf{0.1589 \pm 0.0478}$	**45.83**

†Requires a specific segmentation map not available for SiblingsDB-HQf.

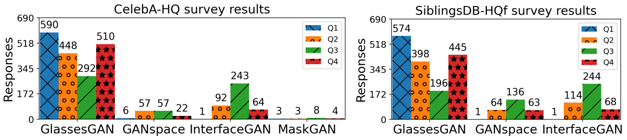

Figure 9. **User-study results.** Raters were asked to choose the method with the best identity preservation (Q1), eyeglasses quality (Q2), realism (Q3), and overall try-on result (Q4). Note that MaskGAN requires a specific segmentation map not available for SiblingsDB-HQf and is, therefore, not included in the right graph.

4.4. Comparison to the State-Of-The-Art

We compare GlassesGAN to three (related) state-of-the-art image-editing techniques utilizing GANs, i.e.: InterFaceGAN [37], MaskGAN [27] and GANSpace [15]. We note that the overall objective of GlassesGAN (i.e., custom design of glasses with visual feedback) is distinct, so **no direct competitors are available** in the literature. We, therefore, select the listed methods as our baselines, as they are able to add glasses (on/off) to facial images and (in some cases) ensure limited amounts of appearance control.

Visual comparison. In Figure 8 we present results with a couple of left-out test images from the CelebA-HQ dataset To ensure a fair comparison, we use publicly released code for the baselines and set the hyperparameters in a way that ensures optimal visual results. Additionally, we select a discrete setting (i.e., appearance of glasses) for GlassesGAN that results in the addition of glasses similar to those produced by the competing methods. As can be seen from the results, all methods generate realistic eyeglasses, but except for GlassesGAN also introduce significant identity changes. While this is a common issue with latent-space based techniques, our framework avoids such problem through the use of blending, which leads to excellent edit locality compared to the baselines. In contrast, the baseline methods introduce undesirable global changes to the facial appearance, as also highlighted by the squared pixel differences in Figure 8.

Quantitative comparison. Next, we perform a quantitative comparison with the state-of-the-art on the designated (left-out) images from the CelebA-HQ and SiblingsDB-HQf datasets. We add glasses to the test images, using a similar procedure as for the visual comparison discussed above. Following established evaluation methodology [27,32], we analyze the results in Table 2 from four different perspectives: (i) through *Mean Square Error* (MSE) scores, computed between the original and edited images, to quantify unwanted pixel-level changes in the edited images, (ii) through *Identity Discrepancy Scores* (IDS), measured with Euclidean distances of the embeddings produced by a pre-trained ArcFace model [4] from the original and edited samples, to capture potential identity changes, introduced by the editing, (iii) through Fréchet Inception Distances (FID) [13] with the original input samples that reflect the

Table 3. **Ablation-study results.** The left table shows results with (w) and without (w/o) image blending, and the right with (w) and without (w/o) subspace tuning.

GlassesGAN Version	CelebA-HQ MSE (↓)	IDS (↓)	FID (↓)
w/o Blending	0.0161	0.6231	80.72
w Blending	**0.0029**	**0.1707**	**26.02**

GlassesGAN Version	CelebA-HQ ERS [in %]
w/o Tuning	29.12%
w Tuning	**6.22%**

realism and quality of the edited images, and (iv) through a user study with $4,704$ responses from 12 human evaluators. For the study, evaluators were shown randomly selected test images and randomly ordered edits from each method and asked to choose the best identity preservation (Q1), quality (Q2), realism (Q3), and overall try-on result (Q4).

From the results in Table 2 and Figure 9, we observe that GlassesGAN leads to significantly lower MSE scores on both datasets, suggesting that the GlassesGAN edits are closest to the originals among all tested methods. Our approach also has substantially less identity drift from the editing process, as shown by the IDS scores that are lower by a factor of 3 compared to the closest competitor and the average user preference of 99% on Q1. Additionally, the edited images generated by GlassesGAN result in the highest perceptual similarity to the original samples among all tested methods, as evidenced by the lowest observed FID scores (see Table 3 for results without blending). Finally, the user survey results in Figure 9 show a general user preference for GlassesGAN over the baseline methods.

4.5. Ablation Studies

We present ablation studies that investigate the impact of (i) image blending and (ii) subspace initialization.

Image blending. The purpose of image blending is to improve the preservation of the subjects' identity and the locality of the edits. Focusing on image identity first, we compare the average IDS scores between the original and edited images with (w) and without (w/o) blending. As we show on the left part of Table 3, blending substantially reduces the average identity discrepancy. Furthermore, it also reduces the MSE scores by more than $5\times$ and the FID scores by more than $3\times$. These results are further supported by the visual results on the left of Figure 10, where blending is again seen to have a beneficial effect on the editing output.

Subspace initialization. The subspace initialization procedure is designed to improve the robustness of glasses manipulations by normalizing the latent-space edits dynamically on a per-sample basis to constrain the edited latent embeddings to the well-defined part of the learned subspace. To quantify the effectiveness of this solution, we develop a performance measure, we refer to as *Edit Robustness Score* (ERS). ERS is defined as the probability that a latent-space edit fails because it is not conducted within the relevant edit space, which in turn leads to editing outputs without glasses. Failed edits with missing glasses are identified with a face parser (S) based on the area of the glasses frames, and the failure probability is estimated on the test images of CelebA-HQ. As can be seen from the right part of Table 3, the subspace initialization helps to reduce ERS scores by a factor close to $5\times$ and makes the editing process significantly more consistent. This can also be seen from the visual example on the right of Figure 10.

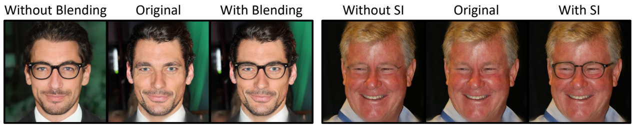

Figure 10. **Visual ablation-study results.** The left images show sample results with (w) and without (w/o) image blending, and the right w and w/o subspace initialization.

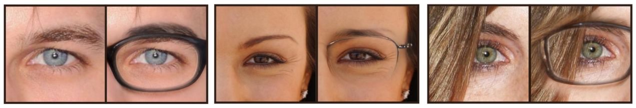

Figure 11. **Illustration of limitations.** Each presented pair shows the original image on the left and the edited one on the right. Parser errors, occlusions, and unusual image characteristics are the main causes of weaker results with a small fraction of the test images.

4.6. Limitations

In Figure 11, we present some limitations of GlassesGAN. Because the framework relies on a face parser S, parser errors may affect the visual quality of the generated results. In the left most example, we see that eyebrows are segmented as part of the frames, leading to changes in appearance. In the middle example, an incorrectly estimated frame area results in improper subspace tuning and poorly visible glasses. In the right, we see that rare occlusions by hair may result in glasses rendered in front of instead of behind the hair. While the visual quality of these examples is still reasonable, such errors are expected to benefit from future advancement in the auxiliary models, e.g., parser S.

5. Conclusion

In this paper, we presented GlassesGAN, a framework for facial image editing that allows the addition of different styles of glasses to input images and continuous editing of their appearance. Extensive experiments over diverse test datasets showed that GlassesGAN yields convincing edits across images with rich appearance variations, while comparing favorably to competing methods. Even though the framework was designed to allow custom creation of glasses, facial editing technology in general may also have *unintended negative social impact* as the modified images could be misused for public shaming, fraud, and manipulating public option. Proper safeguards, therefore, need to be taken when deploying such technology in practice.

References

[1] Pedro Azevedo, Thiago Oliveira Dos Santos, and Edilson De Aguiar. An augmented reality virtual glasses try-on system. In *Symposium on Virtual and Augmented Reality (SVR)*, pages 1–9, 2016. 2

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(4):834–848, 2017. 5

[3] Wen-Huang Cheng, Sijie Song, Chieh-Yun Chen, Shintami Chusnul Hidayati, and Jiaying Liu. Fashion meets computer vision: A survey. *ACM Computing Surveys (CSUR)*, 54(4):1–41, 2021. 1, 2

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 7

[5] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *International Conference on Computer Vision (ICCV)*, pages 9026–9035, 2019. 2

[6] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. C-VTON: Context-driven image-based virtual try-on network. In *IEEE/CVF Winter Applications in Computer Vision (WACV)*, pages 3144–3153, 2022. 1, 2

[7] Zhuming Feng, Fei Jiang, and Ruimin Shen. Virtual glasses try-on based on large pose estimation. In *Procedia Computer Science*, volume 131 of *Recent Advancement in Information and Communication Technology:*, pages 226–233, 2018. 2

[8] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8485–8493, 2021. 2

[9] Wei Gong and Laila Khalid. Aesthetics, personalization and recommendation: A survey on deep learning in fashion. *arXiv preprint arXiv:2101.08301*, 2021. 1, 2

[10] 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 (NeurIPS)*, 2014. 2

[11] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7543–7552, 2018. 1, 2

[12] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing (TIP)*, 28(11):5464–5478, 2019. 2

[13] 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 (NeurIPS)*, 30, 2017. 7

[14] Szu-Hao Huang, Yu-I Yang, and Chih-Hsing Chu. Human-centric design personalization of 3d glasses frame in markerless augmented reality. *Advanced Engineering Informatics*, 26(1):35–45, 2022. 2

[15] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. GANSpace: Discovering interpretable GAN controls. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9841–9850, 2020. 2, 7

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1125–1134, 2017. 2

[17] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *European Conference on Computer Vision (ECCV)*, pages 619–635. Springer, 2020. 2

[18] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. Clothformer: Taming video virtual try-on in all module. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10799–10808, 2022. 2

[19] Andrew Jong, Melody Moh, and Teng-Sheng Moh. Virtual try-on with generative adversarial networks: A taxonomical survey. In *Advancements in Computer Vision Applications in Intelligent Systems and Multimedia Technologies*, pages 76–100. IGI Global, 2020. 2

[20] K. Karhunen. *Zur Spektraltheorie stochastischer Prozesse*. Annales Academiae scientiarum Fennicae. Series A. 1, Mathematica-physica. 1946. 3

[21] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. 5

[22] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:852–863, 2021. 2

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4401–4410, 2019. 2

[24] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 2, 4, 5

[25] Siavash Khodadadeh, Shabnam Ghadar, Saeid Motiian, Wei-An Lin, Ladislau Bölöni, and Ratheesh Kalarot. Latent to latent: A learned mapper for identity preserving editing of multiple face attributes in stylegan-generated images. In *IEEE/CVF Winter Applications in Computer Vision (WACV)*, pages 3184–3192, 2022. 2

[26] Gihyun Kwon and Jong Chul Ye. Diagonal attention and style-based gan for content-style disentanglement in image generation and translation. In *International Conference on Computer Vision (ICCV)*, pages 13980–13989, 2021. 2

[27] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. MaskGAN: Towards diverse and interactive facial image manipulation. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5548–5557, 2020. 2, 7

[28] Kanglin Liu, Gaofeng Cao, Fei Zhou, Bozhi Liu, Jiang Duan, and Guoping Qiu. Towards disentangling latent space for unsupervised semantic face editing. *IEEE Transactions on Image Processing (TIP)*, 31:1475–1489, 2022. 2

[29] Davide Marelli, Simone Bianco, and Gianluigi Ciocca. Faithful fit, markerless, 3d eyeglasses virtual try-on. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *International Conference on Pattern Recognition ICPR: Workshops and Challenges*, pages 460–471, 2021. 2

[30] Arthur Niswar, Ishtiaq Rasool Khan, and Farzam Farbiz. Virtual try-on of eyeglasses using 3d model of the head. In *10th International Conference on Virtual Reality Continuum and Its Applications in Industry (VRCAI)*, pages 435–438, 2011. 2

[31] Gaurav Parmar, Yijun Li, Jingwan Lu, Richard Zhang, Jun-Yan Zhu, and Krishna Kumar Singh. Spatially-adaptive multilayer selection for gan inversion and editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11399–11409, 2022. 2

[32] Martin Pernuš, Vitomir Štruc, and Simon Dobrišek. High Resolution Face Editing with Masked GAN Latent Code Optimization. *IEEE Transactions on Image Processing (TIP), MR*, 2022. 2, 5, 7

[33] Fabio Pizzati, Pietro Cerri, and Raoul de Charette. Comogan: continuous model-guided image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14288–14298, 2021. 2

[34] Ori Press, Tomer Galanti, Sagie Benaim, and Lior Wolf. Emerging disentanglement in auto-encoder based unsupervised image content transfer. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[35] Nathalie Remy, Eveline Speelman, and Steven Swartz. Style that's sustainable: A new fast-fashion formula. Technical report, McKinsey Global Institute, 2016. 1

[36] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9243–9252, 2020. 2

[37] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. InterfaceGAN: Interpreting the disentangled face representation learned by GANs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2, 3, 5, 7

[38] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020. 2

[39] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for StyleGAN image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 2, 5

[40] Tiago F. Vieira, Andrea Bottino, Aldo Laurentini, and Matteo De Simone. Detecting siblings in image pairs. *The Visual Computer*, 30(12):1333–1345, 2014. 5

[41] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11379–11388, 2022. 2

[42] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8798–8807, 2018. 2

[43] Zhengwei Wang, Qi She, and Tomas E Ward. Generative adversarial networks in computer vision: A survey and taxonomy. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021. 2

[44] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12858–12867, 2021. 2

[45] Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022. 2, 3

[46] Yanbo Xu, Yueqin Yin, Liming Jiang, Qianyi Wu, Chengyao Zheng, Chen Change Loy, Bo Dai, and Wayne Wu. Transeditor: Transformer-based dual-space gan for highly controllable facial editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7683–7692, 2022. 2

[47] Guoxing Yang, Nanyi Fei, Mingyu Ding, Guangzhen Liu, Zhiwu Lu, and Tao Xiang. L2m-gan: Learning to manipulate latent space semantics for facial attribute editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2951–2960, 2021. 2, 3

[48] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7850–7859, 2020. 2

[49] Xiaoyun Yuan, Difei Tang, Yebin Liu, Qing Ling, and Lu Fang. Magic glasses: From 2d to 3d. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 27(4):843–854, 2017. 2

[50] Boping Zhang. Augmented reality virtual glasses try-on technology based on iOS platform. *EURASIP Journal on Image and Video Processing*, 2018.1:1–19, 2018. 2

[51] Qian Zhang, Yu Guo, Pierre-Yves Laffont, Tobias Martin, and Markus Gross. A virtual try-on system for prescription eyeglasses. *IEEE Computer Graphics and Applications*, 37(4):84–93, 2017. 2

[52] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. DatasetGAN: Efficient labeled data factory with minimal human effort. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10145–10155, 2021. 5

[53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 2