Barbershop: GAN-based Image Compositing using Segmentation Masks

PEIHAO ZHU, KAUST
RAMEEN ABDAL, KAUST
JOHN FEMIANI, Miami University
PETER WONKA, KAUST

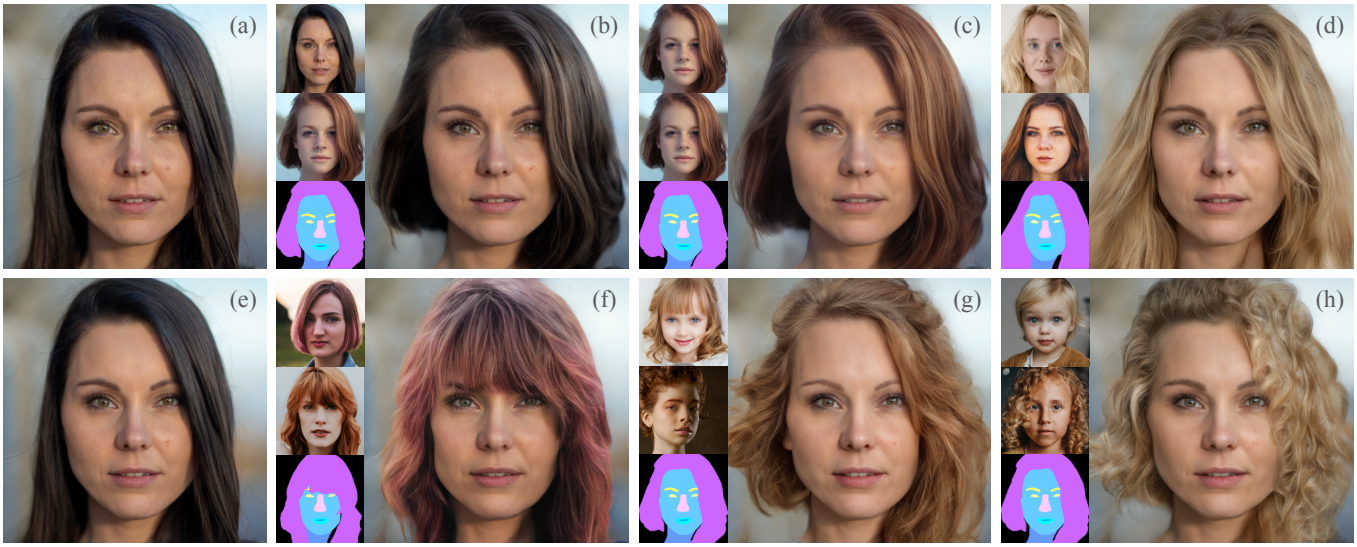

Fig. 1. Hairstyle transfer is accomplished by transferring appearance (fine style attributes) and structure (coarse style attributes) from reference images into a composite image. In each inset the appearance, structure, and target masks for a hairstyle image are shown on the left. Inset (a) is a reference image used for the face and and background, and (e) is a reconstruction using our novel *FS* latent space. In (b) a reference image is used to transfer hair structure, but the hair's appearance is from the original face, and (c) transfers both appearance and structure from a hair reference, in (d) and (f) both structure and appearance attributes are transferred, (g) and (h) use a hair shape that is different from any of the reference images.

Seamlessly blending features from multiple images is extremely challenging because of complex relationships in lighting, geometry, and partial occlusion which cause coupling between different parts of the image. Even though recent work on GANs enables synthesis of realistic hair or faces, it remains difficult to combine them into a single, coherent, and plausible image rather than a disjointed set of image patches. We present a novel solution to image blending, particularly for the problem of hairstyle transfer, based on GAN-inversion. We propose a novel latent space for image blending which is better at preserving detail and encoding spatial information, and propose a new GAN-embedding algorithm which is able to slightly modify images to conform to a common segmentation mask. Our novel representation enables the transfer of the visual properties from multiple reference images including specific details such as moles and wrinkles, and because we do image blending in a latent-space we are able to synthesize images that are coherent. Our approach avoids blending artifacts present in other approaches and finds a globally consistent image. Our results demonstrate a significant improvement over the current state of the art in a user study, with users preferring our blending solution over 95 percent of the time. Project Page and Video: https://zpdesu.github.io/Barbershop.

CCS Concepts: • **Generative Modeling** → **GANs**; • **Image Editing** → *Hairstyle Editing*.

Authors' addresses: Peihao Zhu, peihao.zhu@kaust.edu.sa, KAUST; Rameen Abdal, KAUST, rameen.abdal@kaust.edu.sa; John Femiani, femianjc@miamioh.edu, Miami University; Peter Wonka, pwonka@gmail.com, KAUST.

Additional Key Words and Phrases: Image Compositing, Image Editing, GAN embedding, StyleGAN

1 INTRODUCTION

Due to the rapid improvement of generative adversarial networks (GANs), GAN-based image editing has recently become a widely used tool in desktop applications for professional and social media photo editing tools for casual users. Of particular interest are tools to edit photographs of human faces. In this paper, we propose new tools for image editing by mixing elements from multiple example images in order to make a composite image. Our focus is on the task of hair editing.

Despite the recent success of face editing based on latent space manipulation [Abdal et al. 2019, 2020a; Zhu et al. 2020b], most editing tasks operate on an image by changing global attributes such as *pose*, *expression*, *gender*, or *age*. Another approach to image editing is to select features from reference images and mix them together to form a single, composite image. Examples of composite image editing that have seen recent progress are problems of hair-transfer and face-swapping. These tasks are extremely difficult for a variety of reasons. Chief among them is the fact that the visual properties of different parts of an image are not independent of each-other. The

appearance of hair, for example, is heavily influenced by ambient and reflected light as well as transmitted colors from the underlying face, clothing, and background. The pose of a head influences the appearance of nose, eyes and mouth, and the geometry of a persons head and shoulders influences shadows and the structure of hair. Other challenges include disocclusion of the background, which happens when the hair region shrinks with respect to the background. Disocclusion of the face region can expose new parts of the face, such as ears, forehead, or the jawline. The shape of the hair is influenced by pose and also by the camera intrinsic parameters, and so the pose might have to change to adapt to the hair.

Failure to account for the global consistency of an image will lead to noticeable artifacts - the different regions of the image will appear disjointed even if each part is synthesized with a high level of realism. In order for the composite image to seem plausible, our aim is to make a single coherent composite image that balances the fidelity of each region to the corresponding reference image while also synthesizing an overall convincing and highly realistic image.

Previous methods of hair transfer based on GANs either use a complex pipeline of conditional GAN generators [Tan et al. 2020], each condition module specialized to represent, process, and convert reference inputs with different visual attributes, or make use of the latent space optimization with carefully designed loss and gradient orthogonalization [Saha et al. 2021] to explicitly disentangle hair attributes. While both of these methods show very promising initial results, we found that they could be greatly improved. For example both of them need pretrained inpainting networks to fill holes left over by misaligned hair masks, which may lead to blurry artifacts and unnatural boundaries. We believe that better results can be achieved without an auxiliary inpainting network to fill the holes, as transitions between regions have higher quality if they are synthesized by a single GAN. Also these previous methods do not make use of a semantic alignment step to merge semantic regions from different reference images in latent space, e.g. to align a hair region and a face region from different images.

In this work, we propose Barbershop, a novel optimization method for photo-realistic hairstyle transfer, face swapping, and other composite image editing tasks applied to faces. Our approach uses GAN-inversion to generate high-fidelity reconstructions of reference images. We suggest a novel *FS* latent space which provides coarse control of the spatial locations of features via a *structure tensor* F, as well as fine control of global style attributes via an *appearance code* S. This latent space allows a trade-off between a latent-code's capacity to maintain the spatial locations of features such as wrinkles and moles while also supporting latent code manipulation. We edit the codes to align reference images to a target features locations. This alignment step is a key extension to existing GAN-embedding algorithms. It embeds images while at the same time slightly altering them to conform to a different segmentation mask. Then we find a blended latent code, by mixing reference images in latent space, rather than compositing images in the spatial domain. The result is a latent code of an image. By blending in the new spatially-aware latent space we avoid many of the artifacts of other image compositing approaches.

Our proposed approach is demonstrated in Fig. 1. We are able to transfer only the *shape* of a subject's hair (Fig. 1b). We influence the shape by altering the hair region in a segmentation mask. we can transfer both the shape and the coarse structure (Fig. 1c). We use the term structure to refer to information captured by earlier GAN layers, such as the geometry of the hair strands. For example, the structure encodes the difference between straight, curly, or wavy hair. We can also transfer shape, structure, and detailed appearance (Fig. 1(d,f)). We use the term appearance to describe information encoded in later GAN layers, including hair color, texture, and lighting. Our approach also supports the use of different reference images to be used for structure vs the appearance code as shown in Fig. 1(g,h).

Our main contributions are:

- A novel latent space, called *FS* space, for representing images. The new space is better at preserving details, and is more capable of encoding spatial information.
- A new GAN-embedding algorithm for aligned embedding. Similar to previous work, the algorithm can embed an image to be similar to an input image. In addition, the image is slightly modified to conform to a new segmentation mask.
- A novel image compositing algorithm that can blend multiple images encoded in our new latent space to yield a high quality results.
- We achieve a significant improvement in hair transfer, with our approach being preferred over existing state of the art approaches by over 95% of participants in a user study.

2 RELATED WORK

GAN-based Image Generation. Since their advent, GANs [Goodfellow et al. 2014; Radford et al. 2015] have contributed to a surge in high quality image generation research. Several state-of-the-art GAN networks demonstrate significant improvements in the visual quality and diversity of the samples. Some recent GANs such as ProGAN [Karras et al. 2017], StyleGAN [Karras et al. 2018], StyleGAN2 [Karras et al. 2020b] show the ability of GANs to produce very highly detailed and high fidelity images that are almost indistinguishable from real images. Especially in the domain of human faces, these GAN architectures are able to produce unmatched quality and can then be applied to a downstream task such as image manipulation [Abdal et al. 2019; Shen et al. 2020]. StyleGAN-ada [Karras et al. 2020a] showed that a GAN can be trained on limited data without compromising the generative ability of a GAN. High quality image generation is also attributed to the availability of high quality datasets like FFHQ [Karras et al. 2018], AFHQ [Choi et al. 2020] and LSUN [Yu et al. 2015] objects. Such datasets provide both the quality and diversity to train the GANs and have further contributed to produce realistic applications. On the other hand, BigGAN [Brock et al. 2018] can produce high quality samples using complex datasets like ImageNet [Deng et al. 2009]. Some other notable methods for generative modeling include Variational Autoencoders [Kingma and Welling 2013] (VAEs), PixelCNNs [Salimans et al. 2017], Normalizing Flows [Chen et al. 2018] and Transformer based VAEs [Esser et al. 2020] also have some unique advantages. However, in this work, we focus on StyleGAN2 trained on the FFHQ dataset because it is considered state of the art for face image generation.

Embedding Images into the GAN Latent Space. In order to edit real images, a given image needs to be projected into the GAN latent space. There are broadly two different ways to project/embed images into the latent space of a GAN. The first one is the optimization based approach. Particularly for StyleGAN, I2S [Abdal et al. 2019] demonstrated high quality embeddings into the extended W space, called W+ space, for real image editing. Several followup works [Tewari et al. 2020b; Zhu et al. 2020c] show that the embeddings can be improved by including new regularizers for the optimization. An Improved version of Image2StyleGAN (II2S) [Zhu et al. 2020b] demonstrated that regularization in $P-norm$ space can lead to better embeddings and editing quality. It is also noted that the research in these optimization based approaches with StyleGAN lead to commercial software such as Adobe Photoshop's Neural Filters [Filters [n.d.]]. The second approach in this domain is to use encoder based methods that train an encoder on the latent space. Some notable works [Richardson et al. 2020; Tov et al. 2021] produce high quality image embeddings that can be manipulated. In this work, we propose several technical extensions to build on previous work in image embedding.

Latent Space Manipulation for Image Editing. GAN interpretability and GAN-based image manipulation has been of recent interest to the GAN research community. There are broadly two spaces where semantic manipulation of images is possible: the latent and the activation space. Some notable works in the latent space manipulation domain try to understand the nature of the latent space of the GAN to extract meaningful directions for edits. For instance, GANspace [Härkönen et al. 2020] is able to extract linear directions from the StyleGAN latent space (W space) in an unsupervised fashion using Principal Component Analysis (PCA). Another notable work, StyleRig [Tewari et al. 2020a] learns a mapping between a riggable face model and the StyleGAN latent space. On the other hand, studying the non-linear nature of the StyleGAN latent space, StyleFlow [Abdal et al. 2020b] uses normalizing flows to model the latent space of StyleGAN to produce various sequential edits. Another approach StyleCLIP [Patashnik et al. 2021] uses text information to manipulate the latent space. The other set of papers focus on the layer activations [Bau et al. 2019, 2020] to produce fine grained local edits to an image generated by StyleGAN. Among them are TileGAN [Frühstück et al. 2019], Image2StyleGAN++ [Abdal et al. 2020a], EditStyle [Collins et al. 2020] which try to manipulate the activation maps directly to achieve a desired edit. Recently developed StyleSpace [Wu et al. 2020] studies the style parameters of the channels to produce fine grained edits. StylemapGAN [Kim et al. 2021] on the other hand converts the latent codes into spatial maps that are interpretable and can be used for local editing of an image.

Conditional GANs. One of the main research areas enabling high quality image manipulation is the work on conditional GANs (CGANs) [Mirza and Osindero 2014]. One way to incorporate a user's input for manipulation of images is to condition the generation on another image. Such networks can be trained in either paired [Park et al. 2019; Zhu et al. 2020a] or unpaired fashion [Zhu et al. 2017a,b] using the cycle-consistency losses. One important class of CGANs uses images as conditioning information. Methods such as pix2pix [Isola et al. 2017], BicycleGAN [Zhu et al. 2017b], pix2pixHD [Wang et al. 2018], SPADE [Park et al. 2019], MaskGAN [Fedus et al. 2018], controllable person image synthesis [Men et al. 2020], SEAN [Zhu et al. 2020a] and SofGAN [Chen et al. 2020] are able to produce high quality images given the condition. For instance, these networks can take a segmentation mask as an input and can generate the images consistent with manipulations done to the segmentation masks. Particularly on faces, StarGANs1&2 [Choi et al. 2018, 2020] are able to modify multiple attributes. Other notable works, FaceShop [Portenier et al. 2018], Deep plastic surgery [Yang et al. 2020], Interactive hair and beard synthesis [Olszewski et al. 2020] and SC-FEGAN [Jo and Park 2019] can modify the images using the strokes or scribbles on the semantic regions. For the hairstyle and appearance editing, we identified two notable relevant works. MichiGAN [Tan et al. 2020] demonstrated high quality hair editing using an inpainting network and mask-conditioned SPADE modules to draw new consistent hair. LOHO [Saha et al. 2021] decomposes the hair into perceptual structure, appearance, and style attributes and uses latent space optimization to infill missing hair structure details in latent space using the StyleGAN2 generator. We compare with both these works quantitatively and qualitatively in Sec. 4.2.

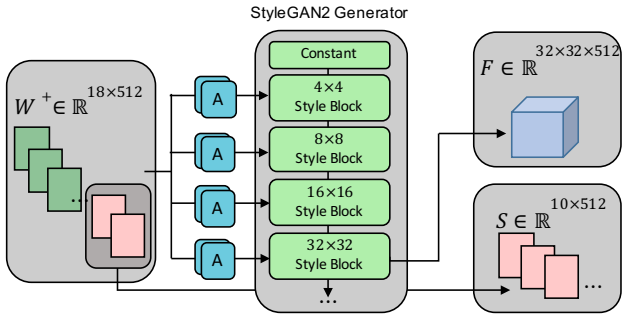

Fig. 2. The latent space. The first eight blocks of the $W+$ code are replaced by the output of the output of the eighth style block to form a structure tensor F, and the remaining parts are used as an appearance code S

.

3 METHOD

3.1 Overview

We create composite images by selecting semantic regions (such as hair, or facial features) from reference images and seamlessly blending them together. To this end, we employ automatic segmentation of reference images and make use of a *target* semantic segmentation mask image $\mathbf{M}$. To perform our most important example edit, hairstyle transfer, one can copy the hairstyle from one image, and use another image for all other semantic categories. More generally, a set of K reference images, $\mathbf{Z}_k$ for $k = 1..K$, are each aligned to the target mask and then blended to form a novel image. The output of our approach is a composite image, $\mathbf{Z}^{\text{blend}}$, in which the region of semantic-category k has the style of reference image $\mathbf{Z}_k$. See Fig. 3 for an overview.

Our approach to image blending is based on StyleGAN [Karras et al. 2020a, 2019, 2020b], and StyleGAN embedding algorithms to

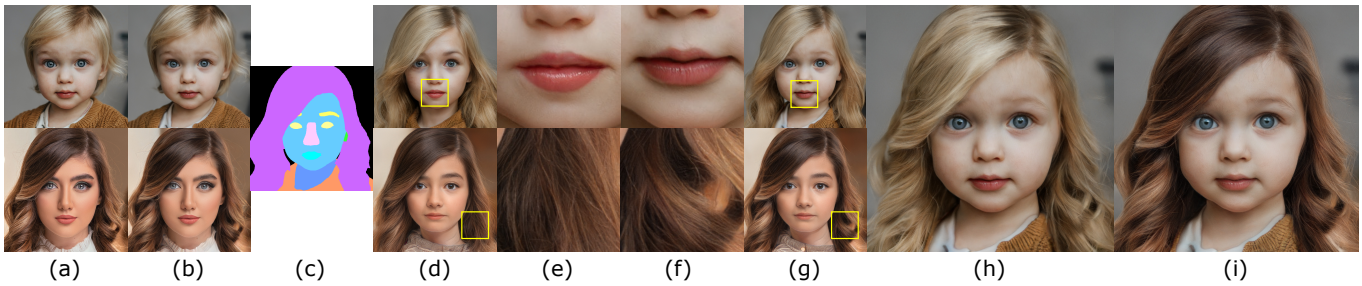

Fig. 3. An overview of the method; (a) reference images for the face (top) and hair (bottom) features, (b) reconstructed images using the *FS* latent space, (c) a target mask, (d) alignment in $W+$ space, (e) a close-up view of the face (top) and hair (bottom) in $W+$ space, (f) close-up views after details are transferred, (g) an entire image with details transferred, (h) the structure tensor is transferred into the blended image, and (i) the appearance code is optimized.

find latent codes for given photographs, e.g. [Abdal et al. 2019]. In particular, we build on the StyleGAN2 architecture [Karras et al. 2020b] and extend the II2S [Zhu et al. 2020b] embedding algorithm. The II2S algorithm uses the inputs of the 18 affine style blocks of StyleGAN2 as a single $W+$ latent code. The $W+$ latent code allows the input of each block to vary separately, but II2S is biased towards latent codes that have a higher probability according to the StyleGAN2 training set. Our approach to image blending finds a latent code for the blended image, which has the benefit of avoiding many of the traditional artifacts of image blending, particularly at the boundaries of the blended regions. However, there is a potential for latent-codes to smooth or elide unusual features of reference images.

In order to increase the capacity of our embedding and capture image details, we embed images using a latent code $\mathbf{C} = (F, S)$ comprised of a *structure tensor*, $F \in \mathcal{R}^{32\times32\times512}$ which replaces the output of style block eight of the StyleGAN2 image synthesis network, and an *appearance code*, $S \in \mathcal{R}^{10\times512}$ that is used as input to the remaining style blocks. This proposed extension of traditional GAN embedding, which we call *FS* space, provides more degrees of freedom to capture individual facial details such as moles. However, it also requires a careful design of latent code manipulations, because it is easier to create artifacts.

Our approach includes the following major steps:

- Reference images are segmented and a *target* segmentation is generated automatically, or optionally the target segmentation is manually edited.
- Individual reference images $\mathbf{Z}_k$ are aligned to the target segmentation and latent codes $\mathbf{C}_k^{\text{align}} = (F_k^{\text{align}}, S_k^{\text{align}})$ are found for the aligned images.
- A combined structure tensor F^{blend} is formed by copying region k of F_k^{align} for each $k = 1...K$.
- Blending weights for the appearance codes S_k^{align} are found so that the appearance code S^{blend} is a mixture of the appearances of the aligned images. The mixture weights are found using a novel masked-appearance loss function.

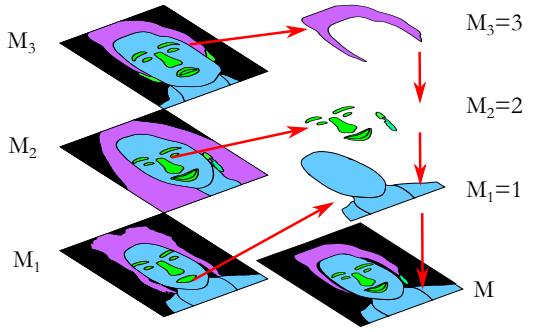

Fig. 4. Generating the target mask. In this example, 19 semantic regions are relabeled to form four semantic categories including background. The label used in the target mask M is the largest index k such that $M_k = k$.

3.2 Initial Segmentation

The first step is to select reference images, (automatically) segment them, and to select regions in the reference images that should be copied to the target image. Let $\mathbf{M}_k = \textsc{Segment}(\mathbf{Z}_k)$ indicate the segmentation of reference image $\mathbf{Z}_k$, where Segment is a segmentation network such as BiSeNET [Yu et al. 2018]. The aim is to form a composite image $\mathbf{Z}^{\text{blend}}$ consistent with a *target* segmentation mask $\mathbf{M}$ so that at locations in the image where $\mathbf{M} = k$, the visual properties of $\mathbf{Z}^{\text{blend}}$ will be transferred from reference images $\mathbf{Z}_k$. The target mask $\mathbf{M}$ is created automatically, however one can also edit the segmentation mask manually to achieve more control over the shapes of each semantic region of the output. In this exposition we will assume that masks are automatically created. Then each pixel target mask $\mathbf{M}(x, y)$ is set to a value k that satisfies the condition that $\mathbf{M}_k(x, y) = k$. If multiple choices of k would satisfy the condition, then the larger k is used. This would happen, for example, if the pixel is covered by *skin* (label 1) in a reference image corresponding to the label *skin*, but also covered by *hair* (label 13) in a reference image corresponding to *hair*, and so the label for *hair* would be chosen. If no choice of k satisfies the condition, then a portion of the target mask will be in-painted using a heuristic method. The process of automatically creating a mask is illustrated in Fig. 4.

3.3 Embedding:

Before blending images, we first align each image to the target mask $\mathbf{M}$. This is important because the appearance of many features such as hair, nose, eyes, and ears depend on the pose of the head as a whole, which introduces a dependency between them. Our approach to aligning the reference images has two parts:

(1) **Reconstruction:** A latent code $\mathbf{C}_k^{\text{rec}}$ is found to reconstruct the input image $\mathbf{Z}_k$.
(2) **Alignment:** A nearby latent code $\mathbf{C}_k^{\text{align}}$ is found that minimizes the cross-entropy between the generated image and the target mask $\mathbf{M}$.

3.3.1 Reconstruction. Given an image $\mathbf{Z}_k$ we aim to find a code $\mathbf{C}_k^{\text{rec}}$ so that $G(\mathbf{C}_k^{\text{rec}})$ reconstructs the image $\mathbf{Z}_k$, where G is the StyleGAN2 image synthesis network. Our approach to finding a reconstruction code $\mathbf{C}_k^{\text{rec}}$ is to initialize it using II2S [Zhu et al. 2020b], which finds a latent code $\mathbf{w}_k^{\text{rec}}$ in the $W+$ latent-space of StyleGAN2. The challenge of any reconstruction algorithm is to find a meaningful trade-off between reconstruction quality and suitability for editing or image compositing. The W latent space of StyleGAN2 has only 512 components, and is not expressive enough to include specific facial details such as moles, wrinkles, or eyelashes. While the latent space is expressive enough to capture generic details, such as wrinkles, it is not possible to encode specific wrinkles in specific locations determined by a reference image. The use of $W+$ space instead of W space improves the expressiveness of the latent space, but it is still not expressive enough to capture specific facial details. One possible approach is noise embedding that leads to embedded images with almost perfect reconstruction, but leads to strong overfitting which manifests itself in image artifacts in downstream editing and compositing tasks. Our idea is to embed into a new latent space, called FS space, that provides better control than $W+$ space without the problems of noise embedding. Similarly to $W+$ embedding, we need to carefully design our compositing operation so that image artifacts do not manifest themselves. The difference between reconstruction in $W+$ vs FS space is shown in Fig. 5, illustrating that key identifying features of a person (such as a facial mole) or important characteristics of a subject's expression (hairstyle, furrows in the brow) are captured in the new latent space.

We capture specific facial details by using a spatially correlated signal as part of our latent code. We use the output of one of the style-blocks of the generator as a spatially-correlated *structure-tensor* F, which replaces the corresponding blocks of the $W+$ latent. The choice of a particular style block is a design decision, however each choice results in a different-sized latent code and in order to keep the exposition concise our discussion will use style-block eight.

The resulting latent code has more capacity than the $W+$ latent-codes, and we use gradient descent initialized by a $W+$-code in order to reconstruct each reference image. We form an initial structure tensor $F_k^{\text{init}} = G_8(\mathbf{w}_k^{\text{rec}})$, and the remaining 10 blocks of $\mathbf{w}_k^{\text{rec}}$ are used to initialize the appearance code S_k^{init}. Then we set $\mathbf{C}_k^{\text{rec}}$ to the nearest local minimum of

$$\mathbf{C}_k^{\text{rec}} = \arg\min_{\mathbf{C}} L_{LPIPS}(\mathbf{C}) + L_F. \tag{1}$$

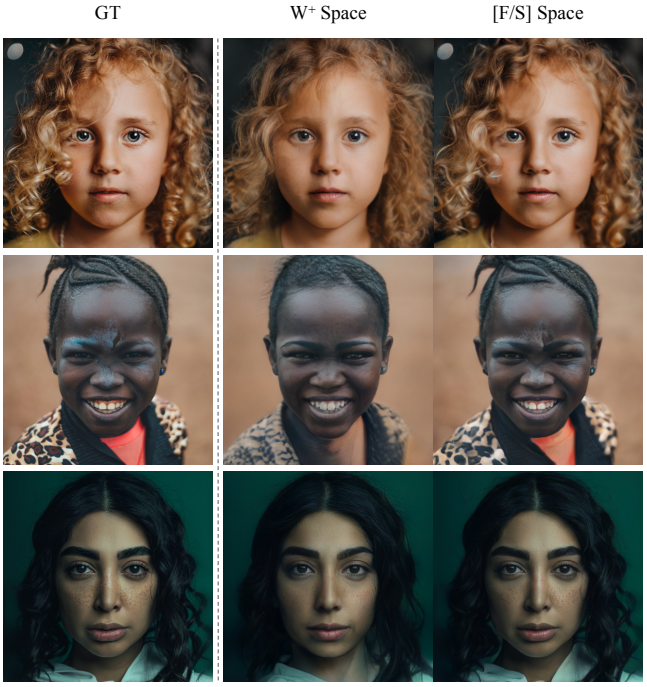

Fig. 5. Reconstruction results on different spaces; (top row) in $W+$ space, structure of the subject's curly hair on the left of the image is lost, and a wisp of hair on her forehead as well as her necklace is removed, but they are preserved in FS space; (middle row) the hair and brow furrows details are important to the expression of the subject, they are not preserved in $W+$ space but they are in FS space; (bottom row) the ground-truth image has freckles, without noise optimization this is not captured in $W+$ space but it is preserved in FS space.

where

$$L_F = \|F - F_k^{\text{init}}\|^2 \tag{2}$$

The term L_F in the loss function (2) encourages solutions in which F remains similar to the activations of a $W+$ code so that the result remains close to the valid region of the StyleGAN2 latent space.

3.3.2 Alignment. We now have each reference image $\mathbf{Z}_k$ encoded as a latent code $\mathbf{C}_k^{\text{rec}}$ consisting of a tensor F_k^{rec} and appearance code S_k^{rec}. While $\mathbf{C}_k^{\text{rec}}$ captures the appearance of the reference image $\mathbf{Z}_k$, the details will not be aligned to the target segmentation. Therefore, we find latent codes $\mathbf{C}_k^{\text{align}}$ that match the target segmentation, and which are nearby $\mathbf{C}_k^{\text{rec}}$. However, directly optimizing $\mathbf{C}_k^{\text{align}}$ is challenging because the details of F_k^{rec} are spatially correlated. Instead we first search for a $W+$ latent code, $\mathbf{w}^{\text{align}}$ for the aligned image and then we transfer details from F_k^{rec} into F_k^{align} where it is safe to do so.

We build on the idea that we can retrieve an image from GAN latent space that conforms to a segmentation mask $\mathbf{M}$ using the cross-entropy loss of the given segmentation mask and a segmentation mask derived from the GAN output using a pre-trained segmentation network. However, here we deal with a specialized version

of this problem. We would like to retrieve a latent representation that conforms to a segmentation mask and that is similar to a given reference image in a given region. Simply initializing with the input representation and then optimizing for a segmentation loss does not work. The image would not be similar enough to the input image. We therefore experimented with a combination of L_2, L_1, and style losses to preserve the content of the reference images and found that only using the style loss produces the best results.

In order to preserve the style between an aligned image $G(\mathbf{w}^{\text{align}})$ and the original image $\mathbf{Z}_k$, we use a masked style-loss. The masked loss described in LOHO [Saha et al. 2021] uses a static mask in order compute the gram matrix of feature activiations only within a specific region, whereas each step of gradient descent in our method produces a new latent code, and leads to a new generated image and segmentation. Therefore the mask used at each step is dynamic. Following [Saha et al. 2021], we base the loss on the gram matrix

$$\mathbf{K}_\ell(\mathbf{Z}) = \gamma_\ell^T \gamma_\ell,$$

where $\gamma_\ell \in \mathcal{R}^{H_\ell W_\ell \times C_\ell}$ is a matrix formed by the activations of layer ℓ of the VGG network. In addition, we define a mask

$$I_k(\mathbf{Z}) = 1\{\textsc{Segment}(\mathbf{Z}) = k\},$$

where $1\{\cdot\}$ is the indicator function, so I_k is an indicator for the region of an image that is of semantic category k. Then the style loss is the magnitude of the difference between the gram matrices of the images generated by a latent code $\mathbf{w}$ and the target image $\mathbf{Z}_k$, and it is evaluated only within semantic region k of each image

$$L_s = \sum_\ell \|\mathbf{K}_\ell(I_k(G(\mathbf{w})) \odot G(\mathbf{w})) - \mathbf{K}_\ell(I_k(\mathbf{Z}_k) \odot \mathbf{Z}_k))\|^2,$$

where the summation is over layers *relu*1_2, *relu*2_2, *relu*3_3, and *relu*4_3 of VGG-16, as was done in LOHO [Saha et al. 2021]. The formulation $I_k(\mathbf{Z}_k) \odot \mathbf{Z}_k$ describes the masking of an image by setting all pixels outside the semantic region k to 0.

In order to find an aligned latent code, we use gradient descent to minimize a loss function which combines the cross-entropy of the segmented image, and the style loss

$$L_{\text{align}}(W) = \textsc{XEnt}(\mathbf{M}, \textsc{Segment}(G(W))) + \lambda_s L_s, \tag{3}$$

where XEnt is the multiclass cross-entropy function. We rely on early-stopping to keep the $\mathbf{w}^{\text{align}}$ latent code nearby the initial reconstruction code $\mathbf{w}^{\text{rec}}$, and λ_s is set to the value recommended by [Saha et al. 2021].

In order transfer the structure and appearance from image $\mathbf{Z}_k$ into F_k, we use binary masks to define safe regions to copy details,

$$\beta_k(x,y) = 1\{\mathbf{M}_k(x,y) = k\},$$
$$\alpha_k(x,y) = 1\{\mathbf{M}(x,y) = k\},$$

where $1\{\cdot\}$ is the indicator function. Let $\beta_{k,\ell}$ denote β_k resized using bicubic-resampling to match the dimensions H_ℓ, W_ℓ of the activations layer ℓ. The mask $\alpha_{k,8} \cdot \beta_{k,8}$ is a region where it is safe to copy structure from the code F_k^{rec} because the semantic classes of the target and reference image are the same. The mask $(1 - \alpha_{k,8} \cdot \beta_{k,8})$ is a region where we must fall-back to $\mathbf{w}_k^{\text{align}}$, which has less capacity to reconstruct detailed features. We use the structure-tensor

$$F_k^{\text{align}} = \alpha_{k,8} \cdot \beta_{k,8} \cdot F_k^{\text{rec}} + (1 - \alpha_{k,8} \cdot \beta_{k,8}) \cdot G_8(\mathbf{w}_k^{\text{align}}),$$

where $G_8(\mathbf{w}^{\text{align}})$ is output of style-block eight of the generator applied to input $\mathbf{w}_k^{\text{align}}$. We now have an aligned latent representation $\mathbf{C}_k^{\text{align}}$ for each reference image k. Next we can composite the final image by blending the structure tensors F_k^{align} and appearance codes S_k^{align} as described in the next two subsections.

3.4 Structure Blending:

In order to create a blended image, we combine the structure tensor elements of $\mathbf{C}_k^{\text{align}}$ using weights α_k to mix the structure tensors, so

$$F^{\text{blend}} = \sum_{k=1}^{K} \alpha_{k,8} \odot F_k^{\text{align}}.$$

The coarse structure of each reference image can be composited simply by combining the regions of each structure tensor, however mixing the appearance codes requires more care.

3.5 Appearance Blending

Our approach to image blending is to find a single style code S^{blend}, which is a mixture of the K different reference codes $S_k, k = 1..K$. To find S^{blend} we optimize a *masked* version of the LPIPS distance function as a loss. Following [Zhang et al. 2018], the distance between an image $\mathbf{Z}$ and $\mathbf{Z}_k$ is

$$L_{\text{LPIPS}} = \sum_{\ell,h,w} \frac{1}{H_\ell W_\ell} \|\boldsymbol{\omega}_\ell \odot (\widehat{y}_\ell(\mathbf{w})_{h,w} - \widehat{y}_\ell(\mathbf{C}_k)_{h,w})\|^2,$$

where $\widehat{y}_\ell$ is the activations of layer ℓ of convnet (VGG) applied to a generated image, and normalized across the channel-dimension, W_ℓ, H_ℓ, and C_ℓ are the shape of a tensor, the vector $\boldsymbol{\omega}_\ell \in \mathcal{R}^{C_\ell}$ has per-channel weights, and the $\odot$ operator indicates elementwise multiplication.

A *masked* version of the loss uses the masks α_k to blend the contributions.

$$L_{\text{masked}} = \sum_{k,\ell,h,w} \frac{\alpha_{k,\ell,h,w}}{H_\ell W_\ell} \cdot \|\boldsymbol{\omega}_\ell \odot (\widehat{y}_\ell(W)_{h,w} - \widehat{y}_\ell(\mathbf{C}_k)_{h,w})\|^2, \tag{4}$$

where $\alpha_{k,\ell}$ is a mask which has been resampled to to match the dimensions of each layer.

Given the K different reference codes $\mathbf{C}_k$, we aim to find a set of k different blending weights $u = \{u_k \in \mathcal{R}^{10\times 512}, k = 1..K\}$. The weights satisfy the constraint that $\sum_k u_k = 1$ and $u_k > 0$. The blended code S^{blend} satisfies

$$S^{\text{blend}} = \sum_k u_k S_k^{\text{align}}$$

so that each element of S^{blend} is a convex combination of the aligned reference codes S_k^{aligned}.

We find $\mathbf{C}_{\text{blend}}$ using projected gradient descent [Landweber 1951]. We initialize the u so that the blended image would be a copy of one of the reference images, and solve for u values that minimize $L_{\text{masked}}(\mathbf{C}_{\text{blend}})$ subject to the constraints that $u_k > 0$ and $\sum_k u_k = \mathbf{1}$.

3.6 Mixing Shape, Structure, And Appearance

We have presented an approach to create composite images using a set of reference images $\mathbf{Z}_k$ in which we transfer the shape of a region, the structure tensor information F_k, and also the appearance information S_k. The LOHO [Saha et al. 2021] approach demonstrated that different reference images can be used for each attribute (shape, structure, and appearance) and our approach is capable of doing the same. We simply use an additional set images $\mathbf{Z}_k^S$ for the appearance information, and we set S_k using the last 10 blocks of the $W+$ code that reconstructs $\mathbf{Z}_k^S$ instead of using the latent code that reconstructs $\mathbf{Z}_k$. The additional images $\mathbf{Z}_k^S$ do not need to be aligned to the target mask. We show example of mixing shape, structure, and appearance in Fig. 1(g,h). The larger structures of the hair (locks of hair, curls) are transferred from the structure reference, and the hair color and micro textures are transferred from the appearance image.

4 RESULTS

In this section, we will show a quantitative and qualitative evaluation of our method. We implemented our algorithm using PyTorch and a single NVIDIA TITAN Xp graphics card. The process of finding an II2S embedding takes 2 minutes per image on average, the optimization in (1) takes 1 minute per image. The resulting codes are saved and reused when creating composite images. For each composite image, we solve equation (3) and then (4) to generate a composite image in an average time of two minutes.

4.1 Dataset

We use a set of 120 high resolution (1024×1024) images from [Zhu et al. 2020b]. From these images, 198 pairs of images were selected for the hairstyle transfer experiments based on the variety of appearances and hair shape. Images are segmented and the *target* segmentation masks are generated automatically.

4.2 Competing methods

We evaluate our method by comparing the following three algorithms: MichiGAN [Tan et al. 2020], LOHO [Saha et al. 2021], and our proposed method.

The authors of LOHO and MichiGAN provide public implementations, which we used in our comparison. However, MichiGAN uses a proprietary inpainting module that the authors could not share. The authors supported our comparison by providing some inpainting results for selected images on request. LOHO also uses a pretrained inpainting network. Based on our analysis, both methods can be improved by using different inpainting networks as proposed in the initial papers. We therefore replaced both inpainting networks by the current state of the art CoModGAN [Zhao et al. 2021] trained on the same dataset as LOHO. All hyperparameters and configuration options were kept at their default values.

Our approach was used to reconstruct images using a fixed number of gradient descent iterations for each step. To solve for $\mathbf{C}_k^{\text{rec}}$ in equation (1) we used 400 iterations, to solve for $\mathbf{C}_k^{\text{align}}$ using (3) we stopped after 100 iterations, and to solve for the blending weights u using (4) we stopped after 600 iterations. **Source code for our method will be made public after an eventual publication of the paper.**

	RMSE↓	PSNR↑	SSIM↑	VGG↓	LPIPS↓	FID↓
Baseline	0.07	23.53	0.83	0.76	0.20	43.99
LOHO	0.10	22.28	0.83	0.71	0.18	56.31
MichiGAN	0.06	26.51	0.88	0.48	0.12	26.82
Ours	**0.03**	**29.91**	**0.90**	**0.38**	**0.06**	**21.21**

Table 1. A comparison of our method to different algorithms using established metrics. Our method achieves the best scores in all metrics.

4.3 Comparison

4.3.1 User Study. We conducted a user study using Amazon's Mechanical Turk to evaluate the hairstyle transfer task. For this task we use the 19-category segmentation from CelebAMask-HQ. A *hairstyle* image was used as the reference for the the corresponding category in CelebAMask-HQ, and an *Identity* image was used for all other semantic categories. We generated composite images using our complete approach and compared the results to LOHO [Saha et al. 2021] and to MichiGAN [Tan et al. 2020]. Users were presented with each image in a random order (ours on the left and the other method on the right, or with ours on the right and the other method on the left). The reference images were also shown at 10% the size of the synthesized images. The user interface allowed participants to zoom in and inspect details of the image, and our instructions encouraged them to do so. Each user was asked to indicate which image combined the face of one image and the hair of another with the highest quality, and fewest artifacts. On average, users spent 90 seconds comparing images before making a selection. We asked 396 participants to compare ours to LOHO, and our approach was selected 378 times (95%) and LOHO was selected 18 times (5%). We asked another 396 participants to compare against MichiGAN, and the results were 381 (96%) ours vs 14 (4%) MichiGAN. The results in both case are statistically significant.

4.3.2 Reconstruction Quality. In this work, we measure the reconstruction quality of an embedding using various established metrics: RMSE, PSNR, SSIM, VGG perceptual similarity [Simonyan and Zisserman 2014], LPIPS perceptual similarity, and the FID [Heusel et al. 2017] score between the input and embedded images. The results are shown in Table 1.

4.4 Ablation Study

We present a qualitative ablation study of the proposed approach for hairstyle transfer. Fig. 9 provides a visual comparison of the results of hairstyle transfer. A *baseline* version of our approach does not include the *FS* latent space and does not do image alignment and is shown in Fig. 9(left column). It does solve for interpolated blending weights to minimize the masked loss function from equation (4), however a mixture of unaligned latent codes does not always result in a plausible image. This is apparent when you compare the face reference image to the synthesized images, which do not faithfully capture the identity of the original subject and in some cases fail to even reconstruct facial features such as eyes when they are partially occluded. The second column of Fig. 9 includes alignment, but it

Fig. 6. Hair style gallery showing different hairstyles applied to a person by varying the hair structure and appearance. Reference images for the hair appearance are shown at the top of each column, Reference images for the hair structure and the target segmentation masks are shown to the left of each row. Also note that in the last two rows, the hair shape is different from the hair shape of the structure reference images.

does not use *FS* space. Without the additional capacity, the reconstructed images are biased towards a generic face image, with more symmetry and less expression, character, and identifying details than the reference images. The subject in row one has an asymmetric expression which is captured by the structure tensor in *FS* space but it is nearly lost using only the $W+$ latent code. Details including the complexion of subject two, and his waxed mustache, are lost without the *FS* embedding. Overall the qualitative examples show that each successive modification to the proposed approach resulted in higher quality composite images.

4.5 Qualitative Results

In this subsection, we discuss various qualitative results that can be achieved using our method.

In Fig. 6 we demonstrate that our framework can generate a large variety of edits. Starting from an initial photograph, a user can manipulate a semantic segmentation mask manually to change semantic regions, copy segmented regions from reference images, copy structure information for semantic regions from reference images, and copy appearance information from reference images. In the figure, we show many results where the shape of the hair, the structure of the hair, and the appearance of the hair is copied from three difference reference images. Together with the source image, that means that information from up to four images contributes to one final blended result image.

In Fig. 7 we demonstrate that our framework can handle edits to other semantic regions different from the hair. We show how individual facial features such as eyes and eyebrows can be transferred from other reference images, how all facial regions can be copied, and how all facial regions as well as the appearance can be transferred from other source images. We can also attain high quality results for such edits. We would like to remark that these edits are generally easier to perform than hair transfer.

In Fig. 8 we show selected examples to illustrate why our method is strongly preferred compared to the state of the art by users in the user study. While previous results gives good results to this very challenging problem, we can still achieve significant improvements in multiple aspects. First, one can carefully investigate the transition regions between hair and either the background and the face to see that previous work often creates hard transitions, too similar to copy and pasting regions directly. Our method is able to better make use of the knowledge encoded in GAN latent space to find semantic transitions between images. Second, other methods can easily create artifacts, due to misalignment in reference images. This manifests itself for example in features, e.g. hair structure, being cut off unnaturally at the hair boundary. Third, our method achieves a better overall integration of global aspects such as lighting. The mismatch in lighting also contributes to lower quality transitions between hair regions and other regions in other methods. By contrast, other methods also have some advantages over our method. Previous work is better in preserving some background pixels by

Fig. 7. Face swapping results achieved by our method. Each example shows three smaller insets on the left: a reference image (top left) from where the components of the face are transferred, an identity image (middle left) and the target segmentation mask (bottom left). We also vary the appearance of the facial components by changing the appearance reference image (bottom right); first row: examples of eye and eyebrow transfer by varying the appearance reference images; second row: examples of eye, eye brows, nose, mouth and teeth transfer by varying the appearance reference images and keeping the complexion the same as the identity image; third row: examples of eye, eye brows, nose, mouth, teeth and complexion transfer by varying the appearance reference images.

design. However, this inherently lowers the quality of the transition regions. We only focus on hair editing for the comparison, because it seems to be by far the most challenging task. This is due to the possible disocclusion of background and face regions, the more challenging semantic blending of boundaries, and the consistency with global aspects such as lighting. Overall, we believe that we propose a significant improvement to the state of the art, as supported by our user study. We also submit all images used in the user study as supplementary materials to enable reviewers to inspect the quality of our results.

4.6 Limitations

Our method also has multiple limitations. Even though we increased the capacity of the latent space, it is difficult to reconstruct under-represented features from the latent space such as jewelry indicated in Fig.10(2,4). Second, issues such as occlusion can produce confusing results. For example, thin wisps of hair which also partially reveal the underlying face are difficult to capture in Fig. 10(3,5). Many details such as the hair structure in Fig. 10(7) are difficult to preserve when aligning embeddings, and when the reference and target segmentation masks do not overlap perfectly the method may fall-back to a smoother structure. Finally, while our method is tolerant of some errors in the segmentation mask input, large geometric distortions cannot be compensated. In Fig. 10(2,7) we show two such examples.

These limitations could be addressed in future work by filtering-out unmatched segmentation as was done by LOHO [Saha et al. 2021], or by geometrically aligning the segmentation masks *before* attempting to transfer the hair shape using regularization to keep the segmentation masks plausible and avoid issues such as Fig. 10(1,7). The details of the structure tensor could be warped to match the target segmentation to avoid issues such as Fig. 10(6). Issues of thin or transparent occlusions are more challenging and may require more capacity or less regularization when finding embeddings.

5 CONCLUSIONS

We introduced Barbershop, a novel framework for GAN-based image editing. A user of our framework can interact with images by manipulating segmentation masks and copying content from different reference images. We presented several important novel components. First, we proposed a new latent space that combines the commonly used $W+$ style code with a structure tensor. The use of the structure tensor makes the latent code more spatially aware and enables us to preserve more facial details during editing. Second, we proposed a new GAN-embedding algorithm for aligned embedding. Similar to previous work, the algorithm can embed an image to be similar to an input image. In addition, the image can be slightly modified to conform to a new segmentation mask. Third, we propose a novel image compositing algorithm that can blend multiple images encoded in our new latent space to yield a high quality result. Our results show significant improvements over the

Fig. 8. Comparison of our framework with two state of the art methods: LOHO and MichiGAN. Our results show improved transitions between hair and other regions, fewer disocclusion artifacts, and a better consistent handling of global aspects such as lighting.

Fig. 9. A qualitative ablation study. We compare a baseline version that blends latent codes without image alignment (left column), a version that used alignment but uses $W+$ rather than FS latent codes (center column), and our complete approach (right column). The reference images for the face, hairstyle, and the target mask are shown top-to-bottom on the left of each row. Each modification improves the fidelity of the composite image.

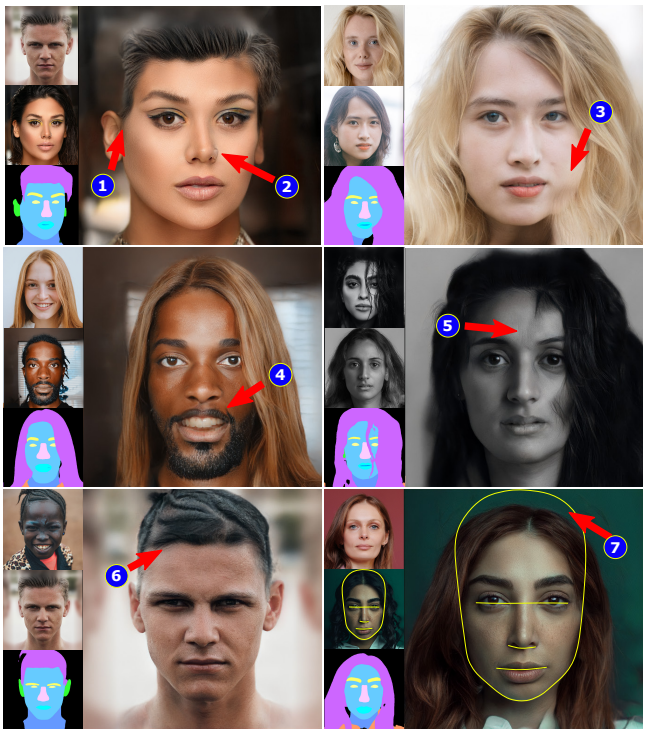

Fig. 10. Failure modes of our approach; (1) misaligned segmentation masks lead to implausible images; (2, 4) the GAN fails to reconstruct the face, replacnig lips with teeth or removing jewelry ; (3,5) overlapping translucent or thin wisps of hair and face pose a challenge; (6) a region of the target mask that is not covered by β_k in the hair image is synthesized with a different structure; (7) combining images taken from different perspectives can produce anatomically unlikely results, the original shape of the head is indicated in yellow.

current state of the art. In a user study, our results are preferred over 95 percent of the time.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2019. Image2stylegan: How to embed images into the stylegan latent space?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4432–4441.

Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020a. Image2stylegan++: How to edit the embedded images?. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8296–8305.

Rameen Abdal, Peihao Zhu, Niloy Mitra, and Peter Wonka. 2020b. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *arXiv e-prints* (2020), arXiv–2008.

David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. 2019. Semantic Photo Manipulation with a Generative Image Prior. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH)* 38, 4 (2019).

David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. 2020. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences* (2020). https://doi.org/10.1073/pnas.1907375117

Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large Scale GAN Training for High Fidelity Natural Image Synthesis. arXiv:1809.11096 [cs.LG]

Anpei Chen, Ruiyang Liu, Ling Xie, and Jingyi Yu. 2020. A Free Viewpoint Portrait Generator with Dynamic Styling. *arXiv preprint arXiv:2007.03780* (2020).

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. 2018. Neural Ordinary Differential Equations. arXiv:1806.07366 [cs.LG]

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Jun 2018). https://doi.org/10.1109/cvpr.2018.00916

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Edo Collins, Raja Bala, Bob Price, and Sabine Süsstrunk. 2020. Editing in Style: Uncovering the Local Semantics of GANs. arXiv:2004.14367 [cs.CV]

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Patrick Esser, Robin Rombach, and Björn Ommer. 2020. Taming Transformers for High-Resolution Image Synthesis. arXiv:2012.09841 [cs.CV]

William Fedus, Ian Goodfellow, and Andrew M. Dai. 2018. MaskGAN: Better Text Generation via Filling in the_____. arXiv:1801.07736 [stat.ML]

Neural Filters. [n.d.]. Adobe Photoshop. https://helpx.adobe.com/photoshop/using/neural-filters.html.

Anna Frühstück, Ibraheem Alhashim, and Peter Wonka. 2019. TileGAN: synthesis of large-scale non-homogeneous textures. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–11.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv:1406.2661 [stat.ML]

Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. 2020. Ganspace: Discovering interpretable gan controls. *arXiv preprint arXiv:2004.02546* (2020).

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*. 6626–6637.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR* (2017).

Youngjoo Jo and Jongyoul Park. 2019. SC-FEGAN: Face Editing Generative Adversarial Network With User's Sketch and Color. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct 2019). https://doi.org/10.1109/iccv.2019.00183

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv:1710.10196 [cs.NE]

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020a. Training Generative Adversarial Networks with Limited Data. In *Proc. NeurIPS*.

Tero Karras, Samuli Laine, and Timo Aila. 2018. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948* (2018).

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020b. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Hyunsu Kim, Yunjey Choi, Junho Kim, Sungjoo Yoo, and Youngjung Uh. 2021. StyleMapGAN: Exploiting Spatial Dimensions of Latent in GAN for Real-time Image Editing. *arXiv preprint arXiv:2104.14754* (2021).

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

Louis Landweber. 1951. An iteration formula for Fredholm integral equations of the first kind. *American journal of mathematics* 73, 3 (1951), 615–624.

Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. 2020. Controllable Person Image Synthesis With Attribute-Decomposed GAN. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2020). https://doi.org/10.1109/cvpr42600.2020.00513

Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. arXiv:1411.1784 [cs.LG]

Kyle Olszewski, Duygu Ceylan, Jun Xing, Jose Echevarria, Zhili Chen, Weikai Chen, and Hao Li. 2020. Intuitive, Interactive Beard and Hair Synthesis With Generative Models. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2020). https://doi.org/10.1109/cvpr42600.2020.00747

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. arXiv:2103.17249 [cs.CV]

Tiziano Portenier, Qiyang Hu, Attila Szabó, Siavash Arjomand Bigdeli, Paolo Favaro, and Matthias Zwicker. 2018. Faceshop. *ACM Transactions on Graphics* 37, 4 (Aug 2018), 1–13. https://doi.org/10.1145/3197517.3201393

Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv:1511.06434 [cs.LG]

Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. 2020. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. *arXiv preprint arXiv:2008.00951* (2020).

Rohit Saha, Brendan Duke, Florian Shkurti, Graham W. Taylor, and Parham Aarabi. 2021. LOHO: Latent Optimization of Hairstyles via Orthogonalization. arXiv:2103.03891 [cs.CV]

Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P. Kingma. 2017. PixelCNN++: A PixelCNN Implementation with Discretized Logistic Mixture Likelihood and Other Modifications. In *ICLR*.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020).

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Zhentao Tan, Menglei Chai, Dongdong Chen, Jing Liao, Qi Chu, Lu Yuan, Sergey Tulyakov, and Nenghai Yu. 2020. MichiGAN. *ACM Transactions on Graphics* 39, 4 (Jul 2020). https://doi.org/10.1145/3386569.3392488

Ayush Tewari, Mohamed Elgharib, Gaurav Bharaj, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. 2020a. Stylerig: Rigging stylegan for 3d control over portrait images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6142–6151.

Ayush Tewari, Mohamed Elgharib, Mallikarjun BR, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zöllhofer, and Christian Theobalt. 2020b. PIE: Portrait Image Embedding for Semantic Control. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)* 39, 6. https://doi.org/10.1145/3414685.3417803

Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. 2021. Designing an Encoder for StyleGAN Image Manipulation. *arXiv preprint arXiv:2102.02766* (2021).

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zongze Wu, Dani Lischinski, and Eli Shechtman. 2020. StyleSpace Analysis: Disentangled Controls for StyleGAN Image Generation. *arXiv preprint arXiv:2011.12799* (2020).

Shuai Yang, Zhangyang Wang, Jiaying Liu, and Zongming Guo. 2020. Deep Plastic Surgery: Robust and Controllable Image Editing with Human-Drawn Sketches. *Lecture Notes in Computer Science* (2020), 601–617. https://doi.org/10.1007/978-3-030-58555-6_36

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. *Lecture Notes in Computer Science* (2018), 334–349. https://doi.org/10.1007/978-3-030-01261-8_20

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365* (2015).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. 2021. Large Scale Image Completion via Co-Modulated Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*.

Jiapeng Zhu, Yujun Shen, Deli Zhao, and Bolei Zhou. 2020c. In-domain gan inversion for real image editing. In *European Conference on Computer Vision*. Springer, 592–608.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017a. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *2017 IEEE International Conference on Computer Vision (ICCV)* (Oct 2017). https://doi.org/10.1109/iccv.2017.244

Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. 2017b. Toward Multimodal Image-to-Image Translation. arXiv:1711.11586 [cs.CV]

Peihao Zhu, Rameen Abdal, Yipeng Qin, John Femiani, and Peter Wonka. 2020b. Improved StyleGAN Embedding: Where are the Good Latents? arXiv:2012.09036 [cs.CV]

Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020a. SEAN: Image Synthesis With Semantic Region-Adaptive Normalization. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Jun 2020). https://doi.org/10.1109/cvpr42600.2020.00515