FacialGAN: Style Transfer and Attribute Manipulation on Synthetic Faces

Ricard Durall[1,2]
ricard.durall.lopez@itwm.fraunhofer.de

Jireh Jam[3]
jireh.jam@stu.mmu.ac.uk

Dominik Strassel[1]
dominik.strassel@itwm.fraunhofer.de

Moi Hoon Yap[3]
m.yap@mmu.ac.uk

Janis Keuper[1,4]
janis.keuper@hs-offenburg.de

[1] Fraunhofer ITWM,
Kaiserslautern, Germany

[2] IWR, University of Heidelberg,
Heidelberg, Germany

[3] Manchester Metropolitan University,
Manchester, United Kingdom

[4] IMLA, Offenburg University,
Offenburg, Germany

Abstract

Facial image manipulation is a generation task where the output face is shifted towards an intended target direction in terms of facial attribute and styles. Recent works have achieved great success in various editing techniques such as style transfer and attribute translation. However, current approaches are either focusing on pure style transfer, or on the translation of predefined sets of attributes with restricted interactivity. To address this issue, we propose FacialGAN, a novel framework enabling simultaneous rich style transfers and interactive facial attributes manipulation. While preserving the identity of a source image, we transfer the diverse styles of a target image to the source image. We then incorporate the geometry information of a segmentation mask to provide a fine-grained manipulation of facial attributes. Finally, a multi-objective learning strategy is introduced to optimize the loss of each specific tasks. Experiments on the CelebA-HQ dataset, with CelebAMask-HQ as semantic mask labels, show our model's capacity in producing visually compelling results in style transfer, attribute manipulation, diversity and face verification. For reproducibility, we provide an interactive open-source tool to perform facial manipulations, and the Pytorch implementation of the model.

1 Introduction

Facial image manipulation is a challenging task that involves generating images whilst preserving the subtle texture of relevant features of the faces. When editing, different levels of structural changes are imposed on key characteristics so that the system steers towards the target direction, with the attempt to synthesize realistic facial images. Generative neural networks have been very successful in this task, due to their ability to extract high-level features, and they have been utilized in different editing tasks. Common editing techniques like image-to-image translation include style transfer [7, 17] and attribute manipulation [6, 24].

| source | reference | mask 1 | synthetic 1 | mask 2 | synthetic 2 |
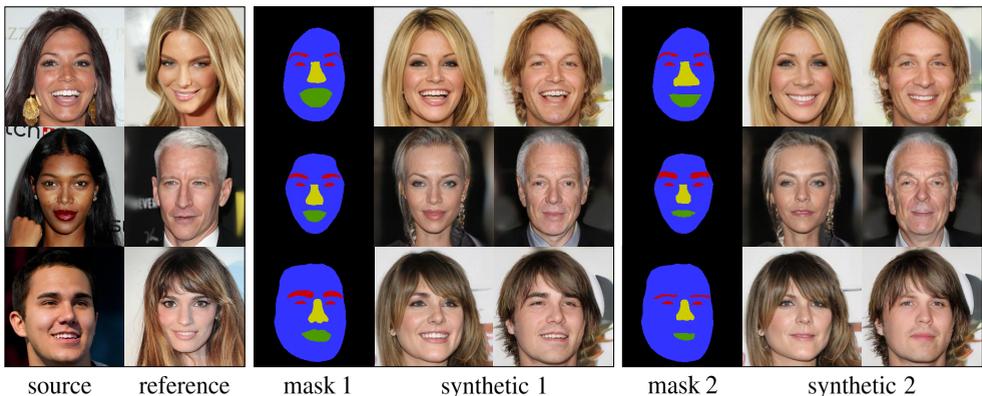
Figure 1: Given a source and a reference image, users can change the mask of the source so that the synthetic result retains the identity from the source, with the style from the reference and follows the customized mask (e.g., modified eyes, eyebrows, nose or mouth).

Style transfer approaches define two or more visual domains, where the goal is to translate from one domain to another. Usually, these domains represent distinguishable properties like gender, hairstyle, and skin colour among others. Ideally, while training these models, one needs to predefine the boundaries of the target domains/styles, as they can be arbitrarily large. To address this problem, [7] introduces an approach, where the styles are controlled either by domain specific encoders or by semantic labels. This structure allows to have an image-to-image translation model that can handle a wide diversity of styles. Despite the remarkable results, this model has two important limiting factors. First, the scalability over multiple domains has a direct impact on the size of the architecture. Additionally, it might require a pre-classification of the data according to the target domain, which is not always a simple task, since some domains cannot be binarised. The second limitation comes from its semantic manipulation nature, i.e., it translates whole images, not allowing local nor pixel-wise manipulations. As a consequence, it cannot have a fine-grained control of the faces, generating less diverse attributes than a geometry manipulation approach.

With facial attribute geometry manipulations, we can achieve rich generation of attributes thanks to its semantic reasoning. It consists of an image generation process, closely bounded to the aforementioned technique, with supervision from a well-defined feature (target attribute), which is modified with consistency preserving realism with the rest of the face. We can categorize attribute manipulation techniques into semantic-level manipulation [6, 23, 26] and geometry-level manipulation [30, 57, 58]. The former is precise and easy to train, however, it does not allow users to interactively manipulate the face images. Recent works on geometry manipulation [11, 24, 54] have achieved notable results using semantic masks as intermediate representations of facial features. While it is true that these methods grant the user with more freedom to manipulate the attributes at will, they are not generating high diversity outputs. [24] overcomes these drawbacks by mimicking user's manipulation via a mask manifold. Nonetheless, this method is limited in terms of style transfer, since it lacks the ability to conduct morphological changes when applying style. A possible solution to mitigate this issue is describing the style via the available attributes, but it is quite restricted.

To address these limitations, we propose a network that learns both styles and semantic attribute translation, thereby ensuring that it can deal with all the aforementioned tasks at

once. Similar to [7], we design a network that extracts and applies diverse styles to generate realistic facial images. Additionally, inspired by [24], we incorporate geometry information via segmentation masks to achieve a fine-grained manipulation of facial attributes, leading to a rich diversity of outputs. We believe that having a functional end-user application is important to both industry and academia, hence, we provide an open-source tool similar to [0], where the user can play around and experiment with our pretrained model[1].

2   Related Work

Generative Adversarial Networks (GANs) [10] have shown impressive results in various computer vision tasks like image generation [3, 21], image-to-image translation [9, 17], inpainting [18, 39] and image segmentation [2, 19]. Among them, facial manipulation tasks have continuously gained attention in recent years due to the high demand of facial editing applications. Facial manipulation can be seen as a multi-domain image-to-image translation problem, where the model works with unique domain features from the face. From a style transfer point of view [4, 7, 17, 25, 26, 29, 33, 41], each domain consists of styles, e.g., hairstyle, makeup and skin colour. From an attribute translation view [5, 6, 13, 24, 30, 31, 34, 35, 36, 37], each domain consists of attributes, e.g., smiling, beard and big nose.

One of the pioneer works on style translation [17] suggested learning to map from a source to a target domain using paired images in a supervised manner. Shortly after, new techniques such as cycle consistency [41] or shared latent space [26] were introduced to remove the need for pairs, reducing in this way the dataset complexity. Concurrently, more advanced topologies were developed, e.g., cascaded refinement [4] or multi-scale [33]. Follow-up works proposed to use disentangled representations based on a domain-invariant content space and a domain-specific attribute space [25], on a maximization of the ratio of the distance between generated images and their latent codes [29], or on a framework that tackled both diversity of generated images and scalability over multiple domains [7].

Attribute translation tasks have also undergone major changes – including semantic-level and geometry-level manipulations. Under a semantic-level umbrella, [36] tried to approach the attribute translation task, by generating swapping attribute-related blocks in the latent space between two images. [31] combined a conditional GAN with an encoder, which allowed manipulating multiple attributes at once. [6] introduced an important breakthrough by employing a single generator to perform multi-domain image translation. [13] also achieved remarkable results with an encoder-decoder architecture, where the attribute information has been treated as a part of the latent representation. For a finer control, [35] primarily took advantage of relative attributes, which described the desired change on selected attributes. As for geometry-level attribute approaches, [37] introduced a new model that used two images of opposite attributes as inputs, to transfer exactly the same type of attributes from one image to another by exchanging certain part of their encodings. A different approach followed [30], where they employed the input layout for modulating the activations in normalization layers through a spatially-adaptive, learned transformation. [5] suggested to decompose facial attributes into multiple semantic components, each corresponding to a specific face region. Similarly, [34] incorporated the influence region of each attribute into the generator, and they combined it with a multi-level patch-wise discriminator structure. Finally, [24] enabled diverse and interactive face manipulation via semantic masks that served as intermediate representations for flexible face manipulations with fidelity preservation.

---

[1]Pytorch implementation is available at the https://github.com/cc-hpc-itwm/FacialGAN

3 Contributions

In this work, we focus on the challenging task of facial image editing. Given an input face image, a target style face image, and a guidance segmentation label mask, our novel framework is able to synthesize an output image that (1) shares a similar style with the target style image, while preserving the input face identity, and (2) follows accurately the semantic mask. To the best of our knowledge, FacialGAN is a pioneer of incorporating both techniques under the same umbrella, leading to a flexible and fine-grained editing control. As a result, our facial system achieves state-of-the-art scores in reference-guided synthesis, improving seminal works such as [7, 24]. We propose a multi-objective training that is able to balance the different components of the architecture. In particular, we introduce a new local segmentation loss to encourage the network to follow the geometry specified in the guidance face mask. Unlike [24] where a complex training strategy generates a supervision signal, our segmentation loss back-propagates informative gradients thanks to its locality characteristics. In other words, it exploits the region of interest, i.e., the target pixel-wise attributes. Overall, our contributions are summarized as follows

- We propose a novel model enabling simultaneous rich style transfers and interactive facial attributes manipulation, while maintaining the identity.

- We introduce an intuitive local segmentation loss that guarantees the pixel-wise attribute control, simplifying the complex global pipelines of [24].

- We assess both qualitatively and quantitatively results on CelebA-HQ dataset. We report state-of-the-art scores on reference-guided generation, surpassing [7, 24].

- We provide an open-source framework which enables the user through an interactive GUI to manipulate the structure of facial features along with transferring styles.

4 Our Approach

4.1 Problem Formulation

Given a source image $\mathbf{x} \in \mathbb{R}^{H \times W \times 3}$, its segmentation label mask $\mathbf{m} \in \mathbb{R}^{H \times W \times C}$, and an arbitrary reference image $\mathbf{y} \in \mathbb{R}^{H \times W \times 3}$, our goal is to train a model that can transfer the style from $\mathbf{y}$ to $\mathbf{x}$, control the gender domain, and being consistent with the geometry constraint of $\mathbf{m}$. Note that C is the category number of the semantic label.

To obtain images with diverse styles and flexible attribute manipulation, we train our approach on a twofold task: (1) to generate domain-specific style vectors from arbitrary images and random noise, and (2) to synthesize realistic faces following the geometry dictated by pseudo-random masks. By enforcing such a behaviour, the model will learn to reflect the style vectors and the changes on the attributes of the masks, producing images with diversity and scalability over multiple domains.

4.2 Model Architecture

The topology of our proposal is depicted in Figure 2. It is composed of a generative network, a style network, a segmentation network and a discriminative network. By combining each of these blocks sequentially, the ensemble model successfully transfers styles and attributes.
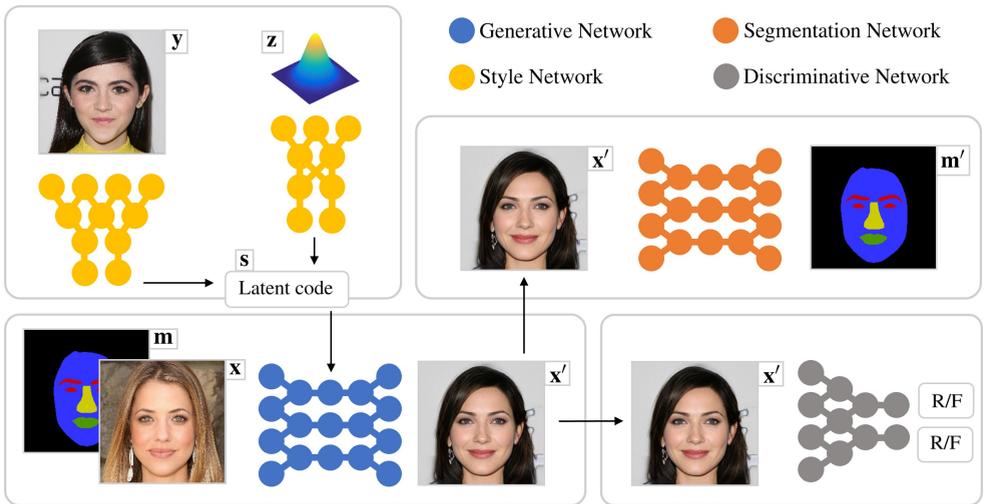
Figure 2: Pipeline of FacialGAN framework. Given an input image **x**, the generator synthesizes a new face **x**′, conditioned on a latent code **s** and a semantic mask **m**. The style network extracts the **s** containing the style. It can be defined either from a reference image **y**, or from random noise **z**. To ensure that the generated face is consistent with the mask, we employ a segmentation network. Finally, the discriminator classifies the output: real (R) or fake (F).

Generative Network. The task of the generator G is to translate an input image **x** into an output image **x**′, following the label mask **m**, while reflecting the style code **s**. Inspired by [9], we use an encoder-decoder topology and randomly mask one of the attributes of **x** so that the network learns to inpaint coherent attributes. Then, we concatenate the masked image with the mask **m**, and feed it into the encoder. For the style transfer, we inject **s** into the decoder using adaptive instance normalization [16]. **Style Network.** The aim of this network is to generate valid style codes **s**. To do that, the network is split into two subnetworks: a mapping network F and an encoder network E. While F generates a style code from random noise **z**, E extracts the style from input images. We adopt the architecture of [7] as a backbone and simplify it for a binary domain to control the gender information. The remaining attributes are manipulated through the mask. **Segmentation Network.** To guarantee a diverse and interactive face manipulation, we need to ensure that G follows the geometry of the mask. Therefore, the segmentation network S generates a control signal that penalizes the generator the moment that the output **x**′ and **m** are not aligned. To achieve that, we feed **x**′ into S and compare the generated output mask **m**′ with the label mask **m** which serve as ground-truth. **Discriminative Network.** The last component is a convolutional discriminator D. However, it behaves slightly different from the vanilla implementation [10], as it takes samples of both real and generated faces and tries to correctly classify them into real and fake based on the gender domain. This discrimination procedure is called "multi-task classification" and it has been successfully employed in previous works [7, 27].

4.3 Multi-Objective Learning

Learning to synthesize realistic and diverse images while transferring styles and manipulating attributes is a complex task. It requires different regularizer terms that focus on specific

tasks. In this work, we mainly use five independent losses in training to achieve our goal.

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{x},\mathbf{m},\mathbf{z}}[\log(1 - D(G(\mathbf{x},\mathbf{m},F(\mathbf{z}))))] \tag{1}$$

The adversarial loss $\mathcal{L}_{\text{adv}}$ [10] is the core element in any GAN-based model. Essentially, it makes the generated images more realistic and assess the control over gender's domain.

$$\mathcal{L}_{\text{sty}} = \mathbb{E}_{\mathbf{x},\mathbf{m},\mathbf{z}}[||\mathbf{s} - E(G(\mathbf{x},\mathbf{m},\mathbf{s}))||_1] \tag{2}$$

The style loss $\mathcal{L}_{\text{sty}}$ [2] is vital to achieve reliable style transfers. It is responsible to enforce the generator to utilize the style codes $\mathbf{s}$, extracted from $F(\mathbf{z})$, by minimizing the distance between them and the style codes, extracted from E when feeding with generated images.

$$\mathcal{L}_{\text{ds}} = -\mathbb{E}_{\mathbf{x},\mathbf{m},\mathbf{z}_1,\mathbf{z}_2}[||G(\mathbf{x},\mathbf{m},F(\mathbf{z}_1)) - G(\mathbf{x},\mathbf{m},F(\mathbf{z}_2))||_1] \tag{3}$$

By maximizing the distance between two generated images with respect to their corresponding latent codes $\mathbf{z}_1$ and $\mathbf{z}_2$, the diverse sensitivity loss $\mathcal{L}_{\text{ds}}$ [29] forces the generator to explore more minor modes and therefore, to produce more diversity.

$$\begin{aligned}\mathcal{L}_{\text{cyc}} = &\mathbb{E}_{\mathbf{x},\mathbf{m},\mathbf{z}}[||\mathbf{x} - G(\mathbf{x}',\mathbf{m},E(\mathbf{x}))||_1] \\ &\text{with } \mathbf{x}' = G(\mathbf{x},\mathbf{m},F(\mathbf{z}))\end{aligned} \tag{4}$$

The cyclic consistency loss $\mathcal{L}_{\text{cyc}}$ [41] guarantees the preservation of the domain invariant characteristics (e.g., pose), while changing its styles faithfully.

$$\mathcal{L}_{\text{seg}} = -\sum_{h,w} m^{h,w,c} \log S(\mathbf{x}^{h,w,c}) + (1 - m^{h,w,c}) \log(1 - S(\mathbf{x}^{h,w,c})) \tag{5}$$

Finally, the local segmentation loss $\mathcal{L}_{\text{seg}}$ is based on binary cross-entropy with the singularity that it works locally. Depending on the manipulated attribute c, $\mathcal{L}_{\text{seg}}$ will evaluate a certain image region (h,w). The goal of this loss is to ensure that the mask rules the attribute geometry of output images. The overall objective can be formulated as

$$\min_{G,F,E,S} \max_{D} \mathcal{L}_{\text{final}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{sty}}\mathcal{L}_{\text{sty}} + \lambda_{\text{ds}}\mathcal{L}_{\text{ds}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cy}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}}, \tag{6}$$

where λ_{adv}, λ_{sty}, λ_{ds}, λ_{cyc} and λ_{seg} are the hyperparameters for each term.

5 Experiments

5.1 Experimental Setup

Baselines Models. We choose state-of-the-art DRIT [25], MSGAN [29], SPADE [30], Star-GANv2 [2] and MaskGAN [24] as our baselines for comparison. DRIT, MSGAN and Star-GANv2 perform latent-guided and reference-guided style transfer. Whereas SPADE and MaskGAN perform geometry-level facial attribute manipulation. **Datasets.** We use CelebA-HQ [20] and CelebAMask-HQ [24] datasets. While CelebA-HQ contains 30,000 high-quality facial images picked from the CelebA [28] dataset, CelebAMask-HQ contains the corresponding semantic segmentation labels separated on 19 classes. For our experiments we resize all images to the size of 256×256, and we create and employ four customized classes – eyes, nose, mouth and skin. **Training Details.** First, we train our segmentation

Method	FID ↓	LPIPS ↑
DRIT [25]	52.1	0.178
MSGAN [29]	33.1	0.389
StarGANv2 [7]	**13.7**	**0.452**
Ours	15.8	0.426

Method	FID ↓	LPIPS ↑
DRIT [25]	53.3	0.311
SPADE [30]	46.2	-
MSGAN [29]	39.6	0.312
MaskGAN [24]	37.1	-
StarGANv2 [7]	23.8	0.388
Ours	**22.8**	**0.415**

Table 1: Distribution-level evaluation on style transfer. (Left) Quantitative comparison on latent-guided synthesis. (Right) Quantitative comparison on reference-guided synthesis.
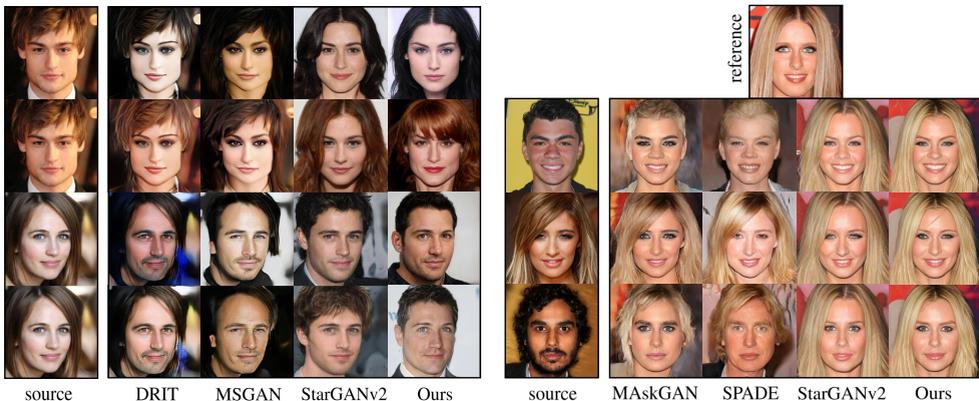
Figure 3: Qualitative comparison of style transfer on image synthesis. (Left) Latent-guided generation using random latent codes. (Right) Reference-guided generation.

model based on [52], for 50 epochs using a batch size of 32, with the default Adam [22] optimizer with learning rate set to 10^{-2}. Then, we train our generative model for 200,000 iterations using a batch size of 8, but in this training we have four Adam optimizers, where we set $\beta_1 = 0$ and $\beta_2 = 0.99$ with learning rates 10^{-4} for G, D and E, and 10^{-6} for F. The losses are all equally weighted except for the segmentation with $\lambda_{\text{seg}} = 2$, and the style diversification where λ_{ds} is linearly decayed to zero over training.

5.2 Evaluation Metrics

Distribution-level Evaluation. To evaluate diversity and visual quality, we use the Fréchet inception distance (FID) [14] and learned perceptual image patch similarity (LPIPS) [40] metrics. **Attribute-level Evaluation.** To evaluate the ability to manipulate target attributes, we train binary facial classifiers for the specific attributes on CelebA beforehand. In particular, we use a ResNet-18 [12] architecture. **Segmentation-level Evaluation.** To evaluate the capacity to generate synthetic images conditioned on the input mask, we train a facial semantic segmentation network on CelebA-HQ. In particular, we use a U-Net [32] architecture that measures the pixel-wise accuracy between the input layout and the predicted parsing results. **Identity-level Evaluation.** In the context of style and attribute facial manipulation, it can be relevant to preserve the identity. Therefore, we employ a pretrained face verification classifier on LFW [15]. In particular, we use ArcFace [8] model with an accuracy of 99.5%.

5.3 Style transfer

We start our experimental evaluation, assessing the style transfer ability of our model from two perspectives: latent-guided synthesis and reference-guided synthesis.

Latent-guided refers to the fact that the system learns to model random noise into valid latent code representations that account for a specific style. Table 1 provides a quantitative comparison of the baseline methods. Our approach provides very competitive results, outperforming most of the models on both FID and LPIPS score, and being very close to [7]. The main reason for these results is the ability of to morphologically change the attributes, resulting in a wide variety of synthetic faces. Our model produces highly diverse results given a single input, having a balanced image quality. Furthermore, we conduct a visual inspection of a few samples. In Figure 3, a qualitative comparison between the different baselines is illustrated. Each column contains the style transfer result from a different random noise input. The top two rows correspond to the results of converting male to female and vice versa in the bottom two rows. We observe that both [7] and our model generate images with a higher visual quality compared to [25] and [29] model. While most of the time [25] synthesizes plausible outcomes, they do not contain morphological changes leading to poorer style transfers. On the other hand, [29] generates results containing more substantial modifications, nonetheless, the method seems to fail to synthesize realistic images.

The second perspective, reference-guided, refers to the fact that the system learns to extract high-level semantics such as hairstyle, makeup, beard and age from the reference images, and to represent it in a latent code. The pose and identity of the source images are preserved. Table 1 shows the quantitative comparison of our method and the baseline methods for reference-guided synthesis. Additionally, we also benchmark the models from [30] and [24]. It should be noted, that these methods are using reference labels to the sources. In this second experiment, our model achieves superior scores in both FID and LPIPS metric compared to the competing models. This implies that our approach produces the most diverse and realistic results while considering the styles of reference images. Figure 3 compares the appearances of FacialGAN with the baseline methods. We observe our approach and [7] have successfully rendered distinctive styles, e.g., hairstyle, makeup and skin colour, while [24] fails at hairstyle translation, and [30] mostly matches only the colours of reference images.

5.4 Attribute transfer

On top of the style transfer capability, our model also allows a fine-grained attributes manipulation based on supervised signals. Those signals come from one hot encoded vector that determine the gender, and from the semantic facial mask that controls the eyes, eyebrows, nose and mouth. Figure 4 shows synthesize images with style and attribute modifications.

We start our analysis by investigating the control over attribute transfers, in particular over the gender. One important difference between some baseline models and ours, is how the gender information is encoded into the generative system. On the one hand, we have [30] and [24] that employ the reference image to determine the gender. In other words, they treat gender as a part of the style information. On the other hand, we have our model and [7] that use a label to set the gender, and therefore, treating it independently of the style. The first column of Table 2 shows the classification accuracy that each baseline achieves on a gender classifier when targeting only male outputs. As one can expect, there is a clear difference between those approaches that use a gender-specific signal, and those who does not. Note that the pretrained classifier has an accuracy of 96.1% which will serve as a ground-truth.

| source | reference | mask 1 | synthetic 1 | mask 2 | synthetic 2 |
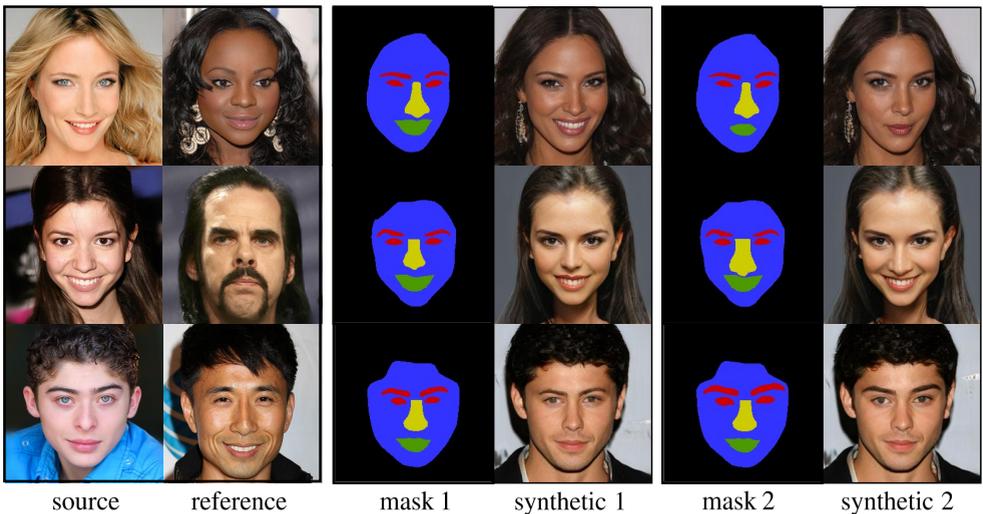
Figure 4: In this picture, we show that our model is able to learn to transform a source image to reflect the style of a given reference image while being consistent with the semantic mask. The source and style reference images appear in the first two columns, whereas the respective transformation masks are given in column 3 and 5. The columns 4 and 6 show the generated images. Each row displays a modified attribute, i.e., mouth, nose and eyes (eyebrows).

Besides gender, our model allows manipulating eyes, eyebrows, nose and mouth independently, and synthesizing images accordingly. As mentioned above, such control comes from the segmentation information of the masks. Our approach is able to react to modifications on the segmentation mask at pixel-level. Hence, it is possible to scale the size of a specific attribute, or even to redraw a new mask completely from scratch. The main limitation in terms of manipulation would arise from the need of realistic customized masks so that they could be translated into realistic faces. We conduct an experiment where we choose *smiling* attribute to compare with previous works [24, 50]. Drawing smiles is a challenging task since not only affects the mouth attribute, but also influences the whole expression of the face, resulting in large geometry variety. To run this evaluation, we manipulate non-smiling masks to be smiling and then, we generate a new set of images. The second column of Table 2 shows how our method achieve very competitive results, outperforming the baselines.

Moreover, we evaluate identity preservation. We first study the effect of style transfer on face recognition, where we measure the accuracy of face recognition of the source and the generated image with style transfer. We achieve an identity accuracy of 89.8%. Further, we apply attribute modification i.e., smiling faces, where we measure the accuracy of face recognition between the source and the generated image with both attribute modification and style transfer. The accuracy in Table 2 shows our method outperforms the baselines with an accuracy of 89.8%. Even with additional modification, our method is able to preserve identity better than the baselines. Having a high attribute transfer accuracy is an important step towards our goal. Nevertheless, this metric might be incomplete as it does not evaluate the precision of the mask, or if the identity is preserved. Hence, we conduct a segmentation study per attribute, see Table 3, where we assess the consistency between the input layout and the predicted parsing results in terms of pixel-wise accuracy.

Method	Male acc.	Smile acc.	Identity acc.
SPADE [30]	54.5	73.8	70.7
MaskGAN [24]	71.7	77.3	76.4
StarGANv2 [7]	**100**	-	-
Ours	**100**	**81.4**	**89.8***
Ground-truth	96.1	92.3	99.5

Table 2: Second and third column show the attribute-level evaluation on male and smiling transfer synthesis accuracy. Fourth, identity-level evaluation after drawing smiles. *Note that our synthetic image also contains style modifications, making more challenge the identity preservation.

Attribute	Ours	GT
Eyes	98.39	98.81
Nose	99.30	99.45
Mouth	98.78	99.06
All	96.40	98.75

Table 3: Segmentation-level evaluation to measure the consistency between the input mask and the predicted parsing results in terms of pixel-wise accuracy.
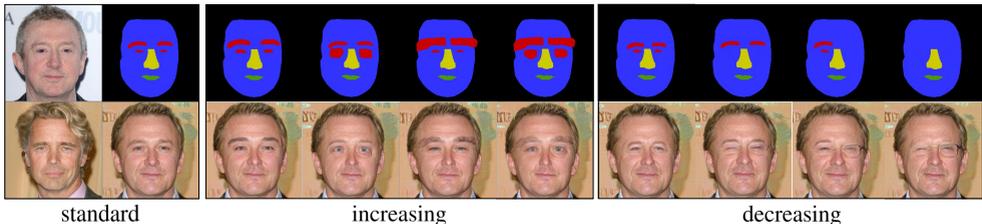
Figure 5: Results on extreme manipulation of eyes (eyebrows) mask.

5.5 Limitations

We run an empirical study to determine under which circumstances our proposal starts to behave erroneously, producing inconsistent outputs. To find such limitations, we constantly increase the size of the mask of the target attributes and evaluate the results. We repeat the same procedure but decreasing the mask size. Figure 5 displays a few examples, where we can see how our model is following the mask if it can be translated into a realistic face. Once the modified mask contains unnatural structures, e.g., no eyes, the network starts to ignore the mask input. The main reason for this behaviour is the effect that the discriminator has over the generator during training, avoiding generating unrealistic faces. A second limitation factor arises from our segmentation loss. In order to generate informative gradients, it needs to work locally. Therefore, we need training data with predefined areas where it will be applied, i.e., the mask of the target attributes. Otherwise, the regions of no interest weaken the learning signal, leading to a loss of control of attribute editing.

6 Conclusions

We propose a novel interactive attribute manipulation and style transfers framework for facial image editing, coined FacialGAN. It learns to extract and to apply diverse style from a reference image while preserving the input face identity, and to incorporate geometry information of a guidance segmentation mask. A multi-objective strategy guarantees the generation of high-quality faces, and a fine-grained manipulation of facial attributes. Experimental results demonstrate that our FacialGAN outperforms state-of-the-art approaches. Finally, we provide an open-source end-user application for the evaluation of our model.

References

[1] Ai face editor. URL https://www.faceapp.com/.

[2] Sergio Benini, Khalil Khan, Riccardo Leonardi, Massimo Mauro, and Pierangelo Migliorati. Fasseg: A face semantic segmentation repository for face image analysis. *Data in brief*, 24, 2019.

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

[4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017.

[5] Ying-Cong Chen, Xiaohui Shen, Zhe Lin, Xin Lu, I Pao, Jiaya Jia, et al. Semantic component decomposition for face attribute manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9859–9867, 2019.

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[7] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[9] Ricard Durall, Franz-Josef Pfreundt, and Janis Keuper. Local facial attribute transfer through inpainting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 95–102. IEEE, 2021.

[10] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

[11] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2019.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[13] 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*, 28(11):5464–5478, 2019.

[14] 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. *arXiv preprint arXiv:1706.08500*, 2017.

[15] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[16] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[18] Jireh Jam, Connah Kendrick, Vincent Drouard, Kevin Walker, Gee-Sern Hsu, and Moi Hoon Yap. R-mnet: A perceptual adversarial network for image inpainting. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2714–2723, 2021.

[19] Mahdi M. Kalayeh, Boqing Gong, and Mubarak Shah. Improving facial attribute prediction using semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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

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

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *NIPS*, 2017.

[24] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020.

[25] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[26] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, pages 700–708, 2017.

[27] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10551–10560, 2019.

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[29] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019.

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019.

[31] Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[33] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[34] Yi Wei, Zhe Gan, Wenbo Li, Siwei Lyu, Ming-Ching Chang, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Maggan: High-resolution face attribute editing with mask-guided generative adversarial network. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[35] Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5914–5922, 2019.

[36] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Dna-gan: Learning disentangled representations from multi-attribute images. *arXiv preprint arXiv:1711.05415*, 2017.

[37] Taihong Xiao, Jiapeng Hong, and Jinwen Ma. Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018.

[38] Weidong Yin, Ziwei Liu, and Chen Change Loy. Instance-level facial attributes transfer with geometry-aware flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9111–9118, 2019.

[39] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4471–4480, 2019.

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

[41] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.