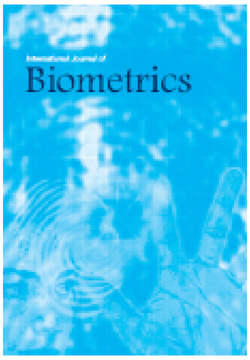


International Journal of Biometrics

ISSN online: 1755-831X - ISSN print: 1755-8301
https://www.inderscience.com/ijbm

Exemplar-based facial attribute manipulation: a review

G. Padmashree, A.K. Karunakar

DOI: 10.1504/IJBM.2024.10054948

Article History:

Received:	16 October 2022
Last revised:	06 January 2023
Accepted:	11 January 2023
Published online:	01 December 2023

Exemplar-based facial attribute manipulation: a review

G. Padmashree and A.K. Karunakar*

Department of Data Science and Computer Applications,
Manipal Institute of Technology,
Manipal Academy of Higher Education,
Manipal, 576104,
Karnataka, India
Email: padmashreeg@gmail.com
Email: karunakar.ak@manipal.edu
*Corresponding author

Abstract: Facial attribute manipulation gained a lot of attention when deep learning algorithms made amazing achievements during the last few years. Facial attribute manipulation is the process of combining or removing desired facial characteristics for a given image. Recently, generative adversarial networks (GANs) and encoder-decoder architecture have been used to tackle this problem, with promising results. We present a comprehensive overview of deep facial attribute analysis from the perspectives of manipulation using exemplars in this study. The model construction approaches, datasets, and performance evaluation measures that are frequently utilised are discussed. Following this, a review of various homogeneous and heterogeneous exemplar-based facial attribute manipulation algorithms is presented in detail. Furthermore, several other facial attribute-related issues and related applications in the real world, are also discussed. Lastly, we go over some of the issues that can arise as well as some interesting future research directions.

Keywords: facial attribute manipulation; image generation; deep learning; generative adversarial networks; GANs; facial attributes; generator; discriminator.

Reference to this paper should be made as follows: Padmashree, G. and Karunakar, A.K. (2024) 'Exemplar-based facial attribute manipulation: a review', *Int. J. Biometrics*, Vol. 16, No. 1, pp.68–111.

Biographical notes: G. Padmashree received his Bachelor's and Master's degrees from the Visvesvaraya Technological University, Karnataka in 2003 and 2012, respectively. She is currently pursuing her PhD research scholar with the Department of Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, India. Her research interests include image processing, artificial intelligence, and machine learning.

A.K. Karunakar received his BSc and MCA from the Karnataka University, Karnataka, India, in 1995 and 1998, respectively, and PhD from the Manipal Academy of Higher Education, Karnataka, in 2009. He is currently a Professor and the Head of the Department of Computer Applications, Manipal

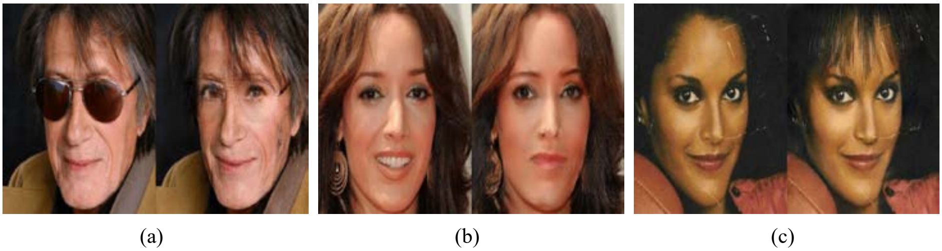
Academy of Higher Education. His research interests include image/video processing and communication, scalable video coding, media-aware network elements, multi-view video coding, scalable video over peer-to-peer networks, error-resilient and concealment for scalable video, stereo vision, and image and video forensics.

1 Introduction

One of the most powerful descriptors for personality attribution is facial features. Computer vision researchers have focused on extracting and exploiting attributes in face recognition. Facial attribute modification aims to change one or more features of a given face image, to generate a fresh face with attributes that we want, and retaining other information like the subject's hair color, gender, etc. The challenge in facial attribute modification is precisely modifying a given image from a source to a target attribute domain while preserving attribute-independent information. Strict geometric constraints and facial attribute correlations must be followed when creating a facial image which is difficult to accomplish. Face attribute editing becomes complicated as a result of these factors. Previous research has concentrated on developing an encoder-decoder architecture for retrieving input image representation and rebuilding it using target attribute vectors as guidance. Generative adversarial networks (GANs) (Goodfellow et al., 2014; Mirza and Osindero, 2014; Chen et al., 2016) and variational autoencoders (VAEs) (Kingma and Welling, 2013; Huang et al., 2018a, 2018b) serve as the backbones for the construction of facial attribute manipulation approaches based on generative models. Figure 1 illustrates some examples of facial attribute modifications. Facial attribute modification can be categorised as model-based which constructs a model without any conditional inputs, and during training learn a set of parameters that solely relate to one attribute extra condition-based methods considers reference images as input conditions that alter multiple attributes instantaneously. On the other hand, extra conditional reference examples exchange specified attributes with the source image during image translation. As a result, attribute transfer using reference images can reveal more detailed characteristics about the source image and provide accurate attribute-modified images (Zhou et al., 2017a; Xiao et al., 2018; Ma et al., 2018). This approach has sparked the interest of many current researchers. Because the generated images include greater facial information and are more lifelike. Figure 2 shows the classification hierarchy of facial attribute manipulation.

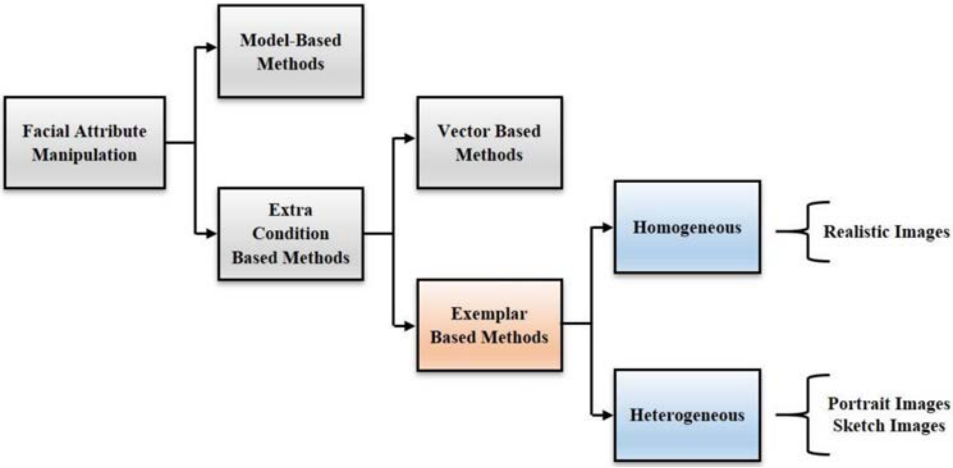
In the literature, there have been many facial attribute manipulation methods used in generating faces with required attributes and retaining the other attribute-independent details. Recently many works have provided promising results for facial attribute transfer. It is observed that model-based methods cannot change multiple attributes in a single training process which led to higher computation time. To overcome this issue, vector-based methods were developed where multiple attributes could be changed at the same time but the main drawback was that they cannot promise that the other details that are not related to the modified qualities will stay the same. Now, face attribute manipulation using the exemplar method has become a research trend as they have the capability of manipulating multiple attributes simultaneously and also preserving the irrelevant attributes.

Figure 1 Examples of facial attribute modifications, (a) removing glasses (b) adding bangs (c) removing smile (see online version for colours)



Source: Xiao et al. (2018)

Figure 2 Classification hierarchy of facial attribute manipulation (see online version for colours)



Facial attribute manipulation can be between either homogeneous or heterogeneous images. Homogeneous images are those in which facial attributes are manipulated between two realistic images and heterogeneous images are those in which facial attribute manipulation happens between realistic and portrait images or sketch images. The latter is one of the latest and most challenging. In this paper, we will be discussing the various face attribute manipulation methods in detail which helps in opening the minds of the researchers to explore this domain.

This paper discusses the various face attribute manipulation methods developed using deep learning which helps in analysing their strength and weakness. Details of different datasets and metrics used for evaluation are presented in Section 2. Section 3 explains the different approaches used in facial attribute manipulation and Section 4 depicts the possible challenges and opportunities that might arise in performing facial attribute manipulation. Finally, the conclusion is provided in Section 5.

2 Model construction, datasets and metrics

2.1 Model construction

2.1.1 Variational autoencoder

A VAE is an architecture that combines an encoder and a decoder and is trained to reduce the reconstruction error between encoded-decoded data and the original data. However, we modify the encoding-decoding procedure somewhat to incorporate some regularisation of the latent space: instead of encoding an input as a single point, we encode it as a distribution throughout the latent space. The input is represented as a latent space distribution. From that distribution, a point from the latent space is sampled. The reconstruction error can be calculated after the sampled point has been decoded as specified in equation (1). Finally, the network propagates the reconstruction error backward. A ‘reconstruction term’ is part of the loss function that is minimised when training a VAE. The Kulback-Leibler (KL) divergence between the returned distribution and a typical Gaussian is the regularisation term. The generator samples the variables *x* parameterised by θ with given latent variables *z* ($p_0(x|z)$).

$$L_{VAE} = \mathbb{E}_{Z\ q_{\phi}(z|x)} \log p_{\theta}(x|z) - D_{KL}(q_0(z|x) \parallel p(z)) \tag{1}$$

2.1.2 Generative adversarial network

GANs are generative models that are trained to generate the input distribution as accurately as possible. Instead of predicting a label’s given features, GAN’s ultimate goal is to predict features given a label. A GAN consists of a generator, *G* that generates new data points from some random uniform distribution *z* obeying a prior noise distribution *z*~*p*(*z*). The goal is to produce a similar type of fake results from inputs, while Discriminator, *D* identifies the fake data produced by *G* from the real data. The generator seeks to persuade the discriminator that the input given by *G* is genuine. Then *G* learns to produce a similar type of training data input similar to min-max game (Goodfellow et al., 2014).

$$\min_{G} \max_{D} L_{GAN} = \mathbb{E}_{x \sim p_{data}(x)} \log(D(x)) + \mathbb{E}_{z \sim p_z(x)} \log(1 - D(G(z))) \tag{2}$$

2.2 Datasets

This section discusses some of the publicly available datasets for facial attribute modification, as indicated in Table 1. Based on the annotations, the datasets may be divided into four categories: multiple attributes, identity, age, and pose and expression. Table 2 lists the various facial features of the face that are examined for manipulation by various datasets.

Table 1 Different datasets available for facial attribute modification

Sl. no.	*Category*	*Dataset*	*Year*	*Available annotations*	*# samples*	*# identities*
1	Multiple attribute	FaceTracer (Kumar et al., 2008)	2008	Facial attribute, Expression	15,000	15,000
2		PubFig (Kumar et al., 2009)	2009	NA	58,797	200
3		YouTube Faces (Wolf et al., 2011)	2011	Identity, pose, expression, illumination	3,245	1,595
4		CelebA (Liu et al., 2015)	2015	40 facial attributes+ 5 landmarks	202,599	10,177
5		CelebA-HQ (Karras et al., 2017)	2017	40 facial attributes+ 5 landmarks	30,000	30,000
6		UMA-AED (Ranjan et al., 2017)	2018	40 facial attributes, illumination, pose, age, skin color	2,800	NA
7		FFHQ (Karras et al., 2019)	2019	Ethnicity, background, age	52,000	NA
8		CelebAMask-HQ (Lee et al., 2020)	2020	19 facial component masks	30,000	30,000
9	Identity	LFW/LFWA (Huang et al., 2007)	2007	Identity, 40 facial attributes	13,233	5,749
10		IJB-A (Klare et al., 2015)	2015	Identity	5,712	500
11		CFP (Sengupta et al., 2016)	2016	Identity, Landmarks	7,000	500
12		IMDb-Face (Wang et al., 2018a)	2018	Identity	1,700,000	59,000
13	Age	FGNET (Fu et al., 2014)	2014	Age	1,002	82
14		CACD (Chen et al., 2014)	2014	Age	163,446	2,000
15		IMDB-WIKI (Rothe et al., 2018)	2015	Age, gender, identity	523,051	20,284
16		Face Aging (Liu et al., 2017)	2017	Age	15,030	15,030
17		AgeDB (Moschoglou et al., 2017)	2017	Age	16,488	568
18		PPB (Buolamwini and Gebru, 2018)	2017	Gender, age	1,270	1,270
19		UTKFace	2017	Age, gender, 68 landmarks	20,000	20,000
20		CLF (Deb et al., 2018)	2018	Age	3,682	919
21		FairFace (Kärkkäinen and Joo, 2019)	2019	Race, gender, age	108,501	108,501
22	Pose and expression	MVF-HQ (Fu et al., 2014)	2009	Pose, expression	120,283	479
23		MultiPIE (Gross et al., 2010)	2010	Pose, expression	750,000	337
24		RaFD (Langner et al., 2010)	2010	Expression, pose, gaze direction	8,040	67
25		FaceWarehouse (Cao et al., 2013)	2013	3D expression	3,000	150

2.2.1 Multiple attribute

This collection of datasets contains several attributes that have been labelled for the manipulation of the facial attribute. *Celeb-Faces Attributes (CelebA)* Dataset is constructed from images selected from Celeb-Faces (Liu et al., 2015). It consists of 10,177 people with 202,599 face images along with 5 landmark locations. Each image has been annotated with 40 attributes. The images in the dataset cover large variations in poses and backgrounds. *CelebA-HQ* (Karras et al., 2017) consists of 30,000 images of high resolutions which are selected from the CelebA Dataset based on CelebA-HQ. Segmentation masks of 19 facial attributes exist for each image of the CelebA dataset which was annotated manually with the size of 512 × 512. To name a few the facial attributes and accessories considered during segmentation are skin, ears, lips, nose, eyes, hair, eyebrows, mouth, necklace, eyeglasses, earrings, and others. *University of Maryland Attribute Evaluation Dataset (UMA-AED)* (Hand et al., 2018a) has been created by considering 40 attributes and HyperFace as face detector (Ranjan et al., 2017). The UMD-AED dataset is utilised as an assessment dataset and contributes to class-imbalance learning for deep face attribute estimation. It's made up of 2800 pictures of people's faces that have been annotated with a subset of 40 CelebA and LFWA traits. Because each attribute has 50 positive and negative samples, not every attribute is labelled in every image. UMD-AED includes a lot of variations, such as various image quality, distinct illuminations and postures, diverse age ranges, and different skin colors. *YouTube Faces Dataset* (with attribute labels). The original YouTube dataset is a database of face videos created to research the topic of unrestricted video face recognition. There are 3,425 recordings in the data collection, with 1,595 different individuals in them. There are approximately 2.15 videos accessible for each subject. The smallest video clip is 48 frames long, while the largest is 6,070 frames long, with an average of 181.3 frames. A total of 620,000 frame images (Wolf et al., 2011) are used for performing face verification. For the problem of video-based face attribute prediction, (Hand et al., 2018b) expanded it further. *Flickr-Faces-HQ Dataset (FFHQ)* (Karras et al., 2019) which is a face database consisting of high-quality images was considered a benchmark for GANs. The image was in PNG format and the resolution of the images was 1,024 × 1,024. The images considered were of varying ethnicity, background, and age along with different accessories such as hats, eyeglasses, etc. and also the face images were aligned. *CelebAMask-HQ* (Lee et al., 2020) is a variant of CelebA-HQ that includes a mask photo for each image in CelebA-HQ. The mask image identifies 19 face components, including the ears, earrings, eyeglasses, eyes, brows, cloth, hair, hat, lip, mouth, nose, neck, necklace, and skin. Similar to these we have FaceTracer (Kumar et al., 2008), and PubFig (Kumar et al., 2009) datasets which contain details of various facial attributes.

2.2.2 Identity

The identity datasets are appropriate for face recognition and identity verification applications since each identity is represented by several distinct face photos. The *Labeled Faces in the Wild (LFW)* database (Huang et al., 2007) consists of 13,233 cropped frontal face images. A total of 5,749 single images of people and 1,680 multiple images of people are collected from online sources. 40 attributes were extracted automatically by (Liu et al., 2015) and five facial landmarks were annotated which leaded to LFWA dataset (Wolf et al., 2010). The *IARPA Janus Benchmark A (IJB-A)* (Klare et al., 2015) database was created to add more obstacles to the face recognition job by gathering facial photos with a wide range of position, illumination, expression, resolution, and occlusion changes. IJB-A is created by gathering an average of 11.4

images and 4.2 videos from 500 Identities, totaling 5,712 images and 2,085 movies. Celebrities in frontal-profile (CFP) (Sengupta et al., 2016), is another face dataset that contains 7,000 face pictures from 500 individuals. The *IMDb-Face* (Wang et al., 2018a) dataset is a large-scale, noise-controlled dataset used for face recognition research. The dataset includes around 1.7 million faces and 59 k identities that were painstakingly cleaned from a total of 2.0 million raw photos. The IMDb website served as the source for all photographs.

2.2.3 Age

To create age-invariant face recognition and verification systems, datasets, in this case, have been labelled with the ages of the people. *FGNet* (Fu et al., 2014) is a dataset for estimating age and recognising faces across ages which is made up of 1,002 photos of 82 persons ranging in age from 0 to 69, with a 45-year age gap. The *Cross-Age Celebrity* (CACD) (Chen et al., 2014) Dataset is a collection of 163,446 photos from 2,000 celebrities gathered from the internet. The names of celebrities and the years (2004–2013) are used as keywords to gather the photographs from search engines. *AgeDB* (Moschoglou et al., 2017) has photos of 16,488 renowned people, including actors/actresses, writers, scientists, politicians, and others. Regarding the identification, age, and gender attributes, each photograph is tagged. With an average of 29 photos per subject, there are 568 different distinct subjects in total. *IMDB-WIKI* (Rothe et al., 2018) is the largest freely available training dataset of face photos labelled with gender and age. The dataset contains a total of 523,051 face photos, including 460,723 from Wikipedia and 20,284 from IMDb's celebrity database. Face Aging (Liu et al., 2017), PPB (Buolamwini and Gebru, 2018), UTKFaces (Zhang et al., 2017), CLF (Deb et al., 2018), and FairFaces (Kärkkäinen and Joo, 2019) are among the other age datasets.

2.2.4 Pose and expression

The *Multi-View Face (MVF-HQ)* (Fu et al., 2014) database contains 120,283 photos with a resolution of 6,000 × 4,000 from 479 different identities with various positions, expressions, and illuminations. MVF-HQ has a substantially larger scale and resolution than publicly accessible high-resolution face manipulation databases. More than 750,000 photos of 337 people are included in the *CMU Multi-PIE* (Gross et al., 2010) face database. Subjects were scanned using 15 view angles and 19 different lighting situations while presenting a variety of facial expressions and high-quality frontal photos. The *Radboud Faces Database (RaFD)* (Langner et al., 2010) is a collection of images of 67 models, comprising Caucasian men and women, Caucasian children, both boys and girls, and Moroccan Dutch men, who are shown with eight different emotional expressions, including anger, disgust, fear, happiness, sadness, surprise, contempt, and neutral. Each emotion was depicted with three different gaze directions, and all photographs were captured from five different camera positions at the same time. *FaceWarehouse* (Cao et al., 2013) is a 3D facial expression database that contains the facial geometry of 150 people of various ages and cultural backgrounds.

Table 2 Facial features of the face that are examined for manipulation by various datasets

Various facial features that are examined for manipulation			
5 o'clock shadow	Colour photo	Middle aged	Senior
Arched eyebrows	Curly hair	Mouth closed	Shiny skin
Asian	Double chin	Mouth slightly open	Sideburns
Attractive	Environment	Mouth wide open	Smiling
Baby	Eyeglasses	Moustache	Soft lighting
Bags under eyes	Face black	Narrow eyes	Straight hair
Bald	Flash	No beard	Strong nose mouth lines
Bangs	Flushed face	No eyewear	Sunglasses
Big lips	Forehead square	Obstructed	Teeth not visible
Big nose	Frowning	Oval face	Teeth visible
Black hair	Fully visible forehead	Pale skin	Wavy hair
Blond hair	Goatee	Pointy nose	Wearing earrings
Blurry	Gray hair	Posed photo	Wearing hat
Brown eyes	Harsh lightening	Receding hairline	Wearing lipstick
Brown hair	Heavy makeup	Rosy cheeks	Wearing necklace
Bushy eyebrows	High cheekbones	Round face	Wearing necktie
Child	Indian	Round jaw	White eyes open
Chubby	Male	Semi obscured forehead	Young

2.3 Metrics

Facial attribute manipulation metrics can be of two types, qualitative where evaluations are performed based on statistical surveys and quantitative refers to how well the facial details and related information are preserved after attribute manipulation. More information on these two metrics is explained below.

2.3.1 Qualitative metrics

In most generative tasks, the most natural way to qualitatively evaluate the quality of generated images is to conduct a statistical survey. Subjects vote on created images with attractive visual quality based on predetermined guidelines, and researchers make results based on the statistics of votes.

For example, (Choi et al., 2018) quantitatively evaluate the performance of generated images in a survey format using 'Amazon Mechanical Turk (AMT)'. Workers are given an input image and told to pick the best-generated images based on subjective reality, attribute transformation quality, and original identity retention. To verify human effort, each individual is presented with a certain series of questions. Zhang et al. (2017) conducted a statistical survey in which volunteers were asked to determine which of their suggested conditional adversarial autoencoder (CAAE) or current works produced the better effect. Sun et al. (2021) ask participants to rate a variety of deep FAM techniques based on subjective reality, quality of transmitted features, and individual trait retention. The average rank (between 1 and 7) of each strategy is then calculated. Lample et al. (2017) evaluate two factors quantitatively: naturalness, which reflects the quality of generated images, and precision, which measures the degree of attribute flipping mirrored in the generation. Wei et al. (2020) evaluated the results by conducting a survey using Amazon Turk. They compared the quality of the images generated using STGAN (Liu et al., 2019) and MagGAN (Wei et al., 2020). Three volunteers were instructed to select the best image which highlights the changed attribute without compromising the quality of the image and also preserving the identity. Chu et al. (2020), Zhang et al. (2020) and Kwak et al. (2020) also performed qualitative analysis by comparing the generated images from the respective models with other state-of-the-art techniques.

2.3.2 Quantitative metrics

Some of the commonly used metrics for comparing the original images and generated images based on distribution difference measures are ‘Fréchet inception distance (FID)’ (Heusel et al., 2017), ‘peak signal to noise ratio (PSNR)’ (Wang et al., 2018b), and ‘structure similarity index (SSIM)’ (Wang et al., 2004). Multi-class classification is performed on the generated images to measure the attribute manipulation accuracy rate (He et al., 2019) using ResNet variants. FID returns the distance between the distributions of original and generated images, PSNR quantifies the quality of the generated images by evaluating the difference between the pixels, and SSIM assesses the structure falsification and the uniqueness distance.

2.3.2.1 Fréchet inception distance

As real-world sample statistics are not taken into account when comparing them to sample statistics from synthetic ones in inception score’s (IS) (Salimans et al., 2016), FID (Heusel et al., 2017) was created. For assessing the effectiveness of GANs, FID is used to measure the quality of generated images which is also known as Wasserstein-2 distance. Using the Inception V3 model, features are extracted from the last pooling layer just before the classification layer. Multivariate Gaussian is obtained by evaluating the mean and covariance of these activations. The Fréchet distance is then used to calculate the distance between these two distributions. It gets its name from the fact that it employs activation’s from the Inception V3 model to summarise each image. Better-quality photos are indicated by lower FIDs; conversely, lower-quality images are indicated by higher scores, and the connection between the two may be linear. Let m_r, m_g, C_r, and C_g represent the feature-wise mean and covariance of the real and generated feature vectors and T_r represents the trace linear algebraic operation. $\|m_r - m_g\|^2$ represents the sum squared mean difference between the two mean vectors. Then the score d^2 is evaluated as given in equation (3). FID has been used as the evaluation metric in Zhu et al. (2019a), Zhang et al. (2019), Ying et al. (2019), Viazovetskyi et al. (2020), Tan et al. (2020), Li et al. (2021), Kowalski et al. (2020), Guo et al. (2019), Esser et al. (2020), Chu et al. (2020), Collins et al. (2020), Abdal et al. (2021), Xiao et al. (2018), Yin et al. (2019), Guo et al. (2019), Lee et al. (2020), Guo et al. (2021), Romero et al. (2021), Dalva et al. (2022), Sun et al. (2022), Phusomsai and Limpiyakorn (2020), Yang et al. (2020b), Deng et al. (2020), Luo et al. (2022), Shi et al. (2022) and Parihar et al. (2022).

$$d^2((m_r, C_r), (m_g, C_g)) = \|m_r - m_g\|^2 + T_r(C_r + C_g + -2(C_r C_g)^{1/2}) \tag{3}$$

2.3.2.2 Peak signal to noise ratio

The mean squared error (MSE) is the most straightforward way to define PSNR (Wang et al., 2018b). MSE is defined as the difference between a noise-free monochromatic image x and its noisy approximation y of size $m \times n$ and given by equation (4). The lower the value of MSE, the lower the error.

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [X(i,j) - Y(i,j)]^2 \tag{4}$$

PSNR is defined in equation (5). As a result, while the PSNR is high, the generated images are of high quality, and when the PSNR is low, the two images are clearly distinguished from one another. Some of the papers which have used this metric are Chen et al. (2020), Guan et al. (2020), Li et al. (2021), Liu et al. (2019), Ning et al. (2021, 2020), Song et al. (2020), Tewari et al. (2020), Wei et al. (2020), Zhang et al. (2020) and Shiri et al. (2019)

$$PSNR = 10 \log_{10} \left(\frac{MAX_X^2}{MSE} \right) \tag{5}$$

2.3.2.3 Structural similarity index

An image quality metric that evaluates the visual impact of three image characteristics: luminance (l), contrast (c), and structure (s) (Wang et al., 2004). It is one of the most widely used metrics in many studies. Let *x* and *y* be the two images being compared and μ_x, μ_y, σ_x, σ_y, and σ_{xy} are the means, standard deviations, and cross-covariance for images *x* and *y*. The mathematical evaluation of luminance, contrast, and structure are defined in equations (6), (7) and (8) respectively. Numerous studies, including Chen et al. (2020), Guan et al. (2020), Li et al. (2021), Liu et al. (2019), Ning et al. (2021, 2020), Song et al. (2020), Tewari et al. (2020), Wang et al. (2021), Shiri et al. (2019) and Deng et al. (2020), use this metric.

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{6}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{7}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{8}$$

C_is and K_is are small constants which is defined as $C_i = (K_i L^2)$ where L is the dynamic range of pixels and $K_i << 1 (1 \leq i \leq 3)$.

Equation (9) gives the multiplicative combination of the three terms resulting in the overall index SSIM. To simplify the expression, if we assume, $\alpha = \beta = \gamma = 1$ and $C3 = C2/2$, we can get, equation (10).

$$SSIM(x,y) = [l(x,y)]^\alpha \cdot [c(x,y)]^\beta \cdot [s(x,y)]^\gamma \tag{9}$$

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{10}$$

The other variations of the structural similarity index are mean-SSIM (MSSIM) and multi-scale-SSIM (MS-SSIM). As the name suggests, MSSIM applies the metric regionally using the circular-symmetric Gaussian Weighing function and then takes the mean overall which is defined in the equation (11). A low-pass filter is used in MS-SSIM to downscale the input images by a factor of 2 after they have been filtered. The original images are indexed as scale 1, and the maximum scale is indexed as scale M, reached after $M - 1$ iterations. Each scale's (*j*) contrast and structural components are determined as $c_j(x,y)$ and $s_j(x,y)$, respectively. Only for the scale M, the luminance component is evaluated as $l_M(x,y)$. Finally, the MS-SSIM is evaluated as defined in equation (12) where α_M, β_j and γ_j adjusts the relative importance of the various components.

$$MSSIM(x,y) = \frac{1}{M}\sum_{j=1}^{M} SSIM(x_j, y_j) \tag{11}$$

$$MS - SSIM(x,y) = [l_M(x,y)]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(x,y)]^{\beta_j} [s_j(x,y)]^{\alpha_j} \tag{12}$$

3 Facial attribute manipulation approaches

Facial attribute manipulation is a process of manipulating the attributes of the source image with the target image which is also called image-to-image transformation. During the transformation, only the facial attributes are taken into consideration rather than identifying photo-realistic images. During the transformation, either a single attribute or multiple attributes can be modified. This section discusses various deep-learning facial attribute manipulation approaches on homogeneous and heterogeneous images. Homogeneous images are those where the source and the target images belong to the same category, i.e., both images are considered to be realistic images. Whereas heterogeneous images are those where the source and target images belong to different categories. For example, the source image can be either a sketch or a portrait image, and the target image would be a realistic image. Tables 3 and 4 provides a summary of the various homogeneous and heterogeneous facial attribute manipulation approaches, advantages, disadvantages, and different datasets used for evaluating their performance of them have been discussed.

3.1 Facial attribute manipulation approaches using homogeneous images

The first example (exemplar)-based facial attribute manipulation method GENEGAN was proposed by Zhou et al. (2017a) as shown in Figure 3. During the manipulation process, the attribute-appropriate information from the source image to the target image is transferred and the attribute-inappropriate information is preserved. The main drawback of this approach is that only one attribute can be altered by one manipulation process.

Table 3 Overview of homogeneous facial attribute manipulation approaches

Methods	*Description*	*Datasets*
Gene-GAN (Zhou et al., 2017a)	Recombing the latent representation information of two paired attribute images for swapping specific attributes.	CelebA (Liu et al., 2015)
ELEGANT (Xiao et al., 2018)	Exchanging latent encoding with GAN for transferring multiple face attributes (ELEGANT) and doing image generation by exemplars as well as producing high-quality generated images.	CelebA (Liu et al., 2015)
Instance-level facial attributes transfer with geometry-aware flow (Yin et al., 2019)	Automatically learns an attribute transfer module and an attribute removal module and jointly operates in a cycle-consistency manner to learn from abundant unpaired data.	CelebA (Liu et al., 2015) and CelebA-HQ (Karras et al., 2017)
MulGAN (Guo et al., 2019)	Attribute labels constraint are applied to the predefined region of the latent feature space and an attribute classification loss is employed.	CelebA (Liu et al., 2015)
FaceShifter (Li et al., 2019)	Attribute labels constraint is applied to the predefined region of the latent feature space and an attribute classification loss is employed.	CelebA-HQ (Karras et al., 2017), FFHQ (Karras et al., 2019) and VGGFace (Parkhi et al., 2015)
MaskGAN (Lee et al., 2020)	Dense mapping network (DMN) learns style mapping between a free-form user-modified mask and a target image, enabling diverse generation results. Editing behaviour simulated training (EBST) models the user editing behaviour on the source mask.	AffectNet (Mollahosseini et al., 2017)
Facial expression manipulation (Wang et al., 2021)	U-Net-based generator with multi-attention gate for facial expression manipulation.	AffectNet (Mollahosseini et al., 2017)
STD-GAN (Guo et al., 2021)	Instance-level facial attribute transfer with style extracting. Weakly supervised attribute style learning with only binary annotations.	CelebA (Liu et al., 2015)
Facial expression manipulation (Ling et al., 2020)	Conditional GAN model employed a multi-level attention mechanism that helped in expression manipulation and identity preservation.	CelebA-HQ (Karras et al., 2017) and CelebAMask-HQ
Mask-adversarial autoencoder (Sun et al., 2021)	VAE-GAN framework modifies a minimum number of pixels in the feature maps of an encoder and allows changing the attribute strength continuously without hindering global information.	CelebA (Liu et al., 2015)
SMILE (Romero et al., 2021)	A multi-attribute image-to-image transformation method for both fine-grained and more global attributes in the semantic space for both random and exemplar-guided synthesis.	CelebA-HQ (Karras et al., 2017) and CelebA-Mask FFHQ (Karras et al., 2019)

Table 3 Overview of homogeneous facial attribute manipulation approaches (continued)

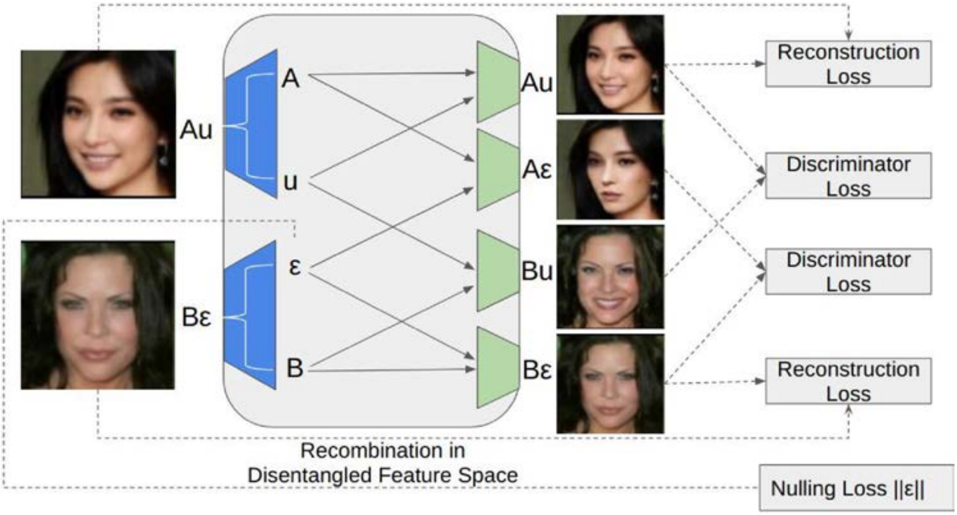
Methods	*Description*	*Datasets*
VecGAN (Dalva et al., 2022)	An image-to-image translation system for modifying facial attributes by factoring latent space.	CelebA-HQ (Karras et al., 2017)
PattGAN (Sun et al., 2022)	The disentangled representation of face attributes from binary attribute labels utilises the disentangled representation of attributes to assist facial attribute editing.	CelebA (Liu et al., 2015)
Large-pose facial makeup (Li and Tu, 2023)	Makeup transfer approach based on generative adversarial networks (GAN) by adopting CycleGAN.	Makeup transfer (Li et al., 2018)
3D GAN inversion with pose optimisation (Ko et al., 2023)	A 3D-GAN inversion technique that incrementally improves an image's latent code and 3D camera posture.	CelebA-HQ (Karras et al., 2017) and FFHQ (Karras et al., 2019)
FastSwap (Yoo et al., 2023)	A simple one-stage system using a triple adaptive normalisation (TAN) block to maintain the identity, pose, and attributes of the inputs.	VoxCeleb2 (Chung et al., 2018)
IA-FaceS (Huang et al., 2023)	A bidirectional approach for disentangled face attribute modification in addition to flexible, controllable component editing.	CelebA-HQ (Karras et al., 2017) and FFHQ (Karras et al., 2019)
SC-GAN (Li et al., 2023b)	A generative adversarial network based on subspace clustering.	CelebA (Liu et al., 2015)
DyStyle (Li et al., 2023a)	An approach to execute nonlinear and adaptive manipulation of latent codes for flexible and precise attribute control.	FFHQ (Karras et al., 2019) and MetFace (Karras et al., 2020)

The ELEGANT model was suggested by Xiao et al. (2018) where multiple facial attributes could be modified in one manipulation process as the encodings of various attributes are available in latent space. Residual learning (He et al., 2016) was incorporated for training the images and multi-scale discriminators (Shen and Liu, 2017) were used for improving the quality of the images. The architecture of the proposed model has been depicted in Figure 4.

Yin et al. (2019) presented an instance-level facial attribute transfer using geometry-aware flow as shown in Figure 5. Facial landmarks are considered geometric guidance for learning the flows automatically. The main advantage of this method is that even though there are huge translation gaps between the poses of the source and target images, the flow can manage strongly. Also, these flows can be applied directly to high-resolution images as they are invariant to scale. Additionally to this, to resolve the appearance gaps, and enhance the quality of the attribute-manipulated image, a refinement sub-network is designed. Finally, attribute transfer and removal modules are designed such that features are learned on ample unpaired facial images.

For transferring multiple facial attributes simultaneously, Guo et al. (2019) developed a novel deep learning method using an encoder-decoder generative network as depicted in Figure 6. The main three components of this model are the generator which is responsible for editing the facial attributes, the attribute classifier is responsible for extracting the attribute-related details from the images and the discriminator is responsible for the generation of photo-realistic images. The network partitions the latent encodings into attribute-relevant and irrelevant along channel dimensions. The attribute-relevant encodings are again partitioned to different attributes where each partition information maps to individual attributes. The attribute labels constraint is then applied directly to the predefined attribute-related blocks in the latent feature space. Before being transmitted to the decoder, binary attribute labels are used to filter predefined attribute-related blocks which own the original attribute label information. Then the classifier makes the model learn the attribute information from the images which helps in the generation of images. Finally, downsampling is applied to reduce the loss and improve the quality of the images generated.

Figure 3 Architecture of GENEGAN (see online version for colours)

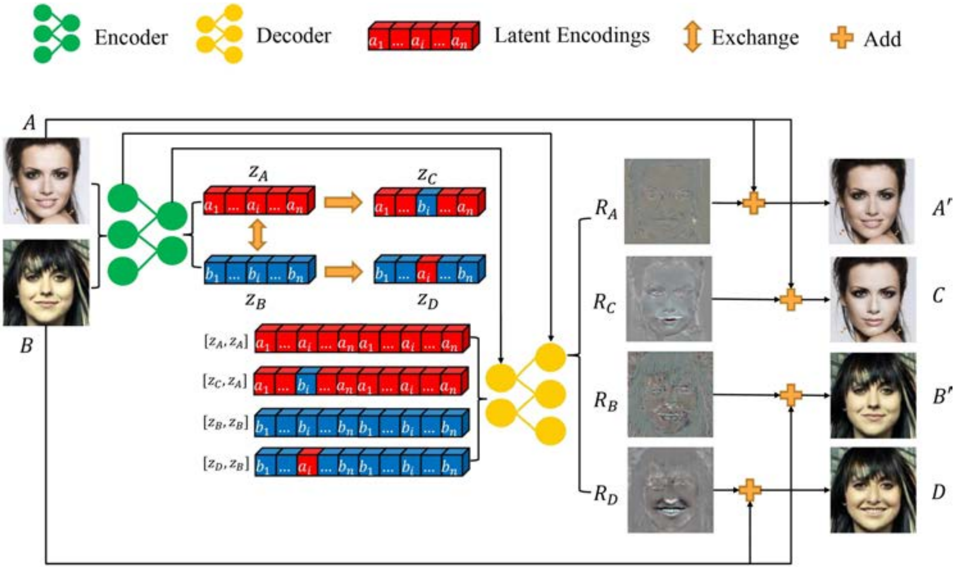


Source: Zhou et al. (2017a)

For extracting multi-level target face attributes, Li et al. (2019) propose an attributes encoder and a generator model using adaptive attentional denormalisation (AAD) layers to for integrating the identity and the attributes. Also, heuristic error acknowledging refinement of network (HEAR-Net), occlusion challenge has been taken care of by recovering anomaly regions in a self-supervised way without any manual annotations. The architecture of the proposed model has been depicted in Figure 7.

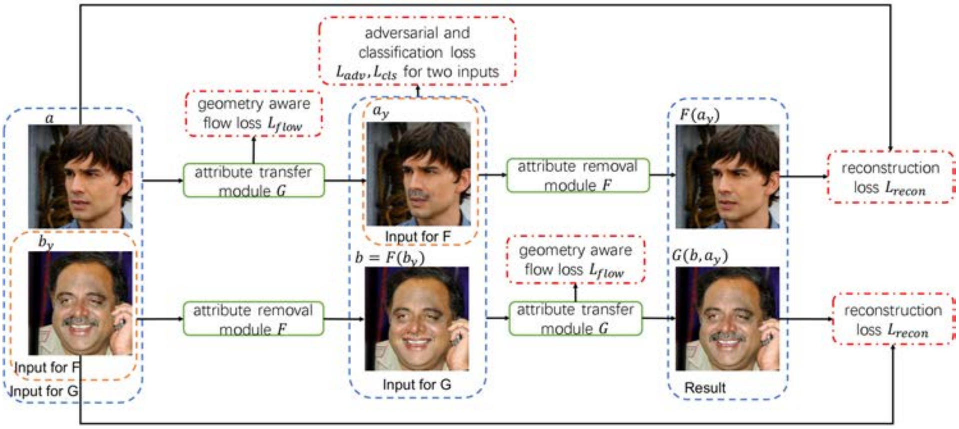
In most of the facial attribute manipulation methods, attribute manipulation could be done only on a predefined set of attributes and there was no facility for interactive attribute manipulation. One of the main findings proposed by Lee et al. (2020) is that in MaskGAN semantic masks are a good intermediate representation for flexible face alteration while maintaining integrity. It consists of two components, dense mapping network (DMN) learns how to translate an unrestricted user-modified mask to a target image, allowing for a wide range of generation results, and editing behaviour simulated training (EBST) models user editing behaviour, making the whole architecture more resistant to varied manipulated inputs. Attribute transfer and style copy are the main two aspects that have been evaluated and have shown better performance. The main drawback was that during face attribute manipulation texture related details could not be controlled. The framework of the proposed model is shown in Figure 8.

Figure 4 Model of ELEGANT for multiple facial attribute manipulation (see online version for colours)



Source: Xiao et al. (2018)

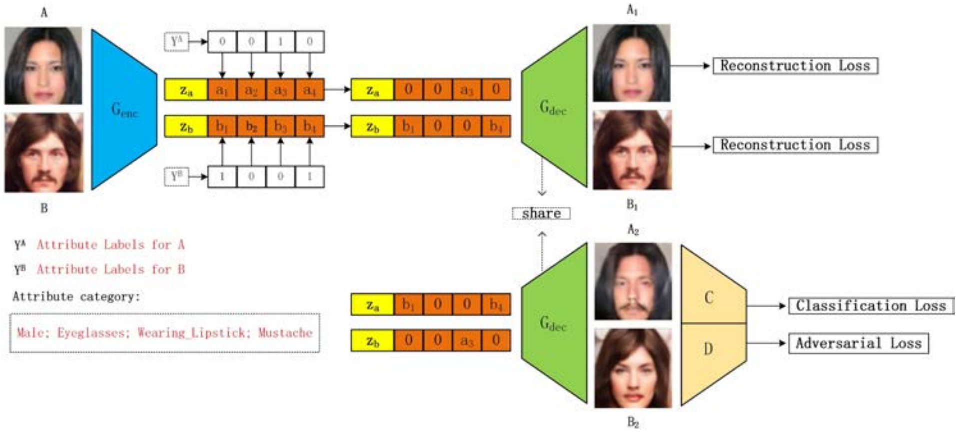
Figure 5 Architecture of instance-level attribute transfer for facial attribute manipulation (see online version for colours)


adversarial and classification loss L_{adv}, L_{cls} for two inputs
geometry aware flow loss L_{flow}
attribute transfer module G
attribute removal module F
reconstruction loss L_{recon}
a
b_y
a_y
$b = F(b_y)$
$F(a_y)$
$G(b, a_y)$
Input for F
Input for G
Result

Source: Yin et al. (2019)

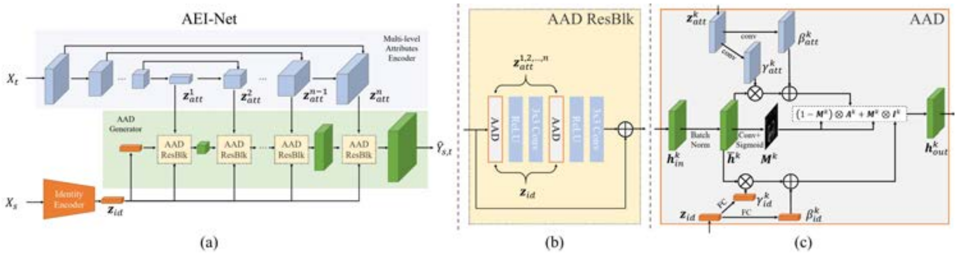
Ling et al. (2020) demonstrated a U-Net-based architecture for facial expression editing that employs a multi-scale fusion mechanism based on relative action units and primarily addresses facial muscle movements. The model was evaluated using both network-based and human-based approaches. Human-based evaluation metrics included expression fulfillment, relative realism, and identity-preserving ability, whereas network-based evaluation metrics included IS, average content distance (ACD), and expression distance (ED). The framework of the proposed model is shown in Figure 9.

Figure 6 MulGAN encoder-decoder generative network for facial attribute manipulation (see online version for colours)



Source: Guo et al. (2019)

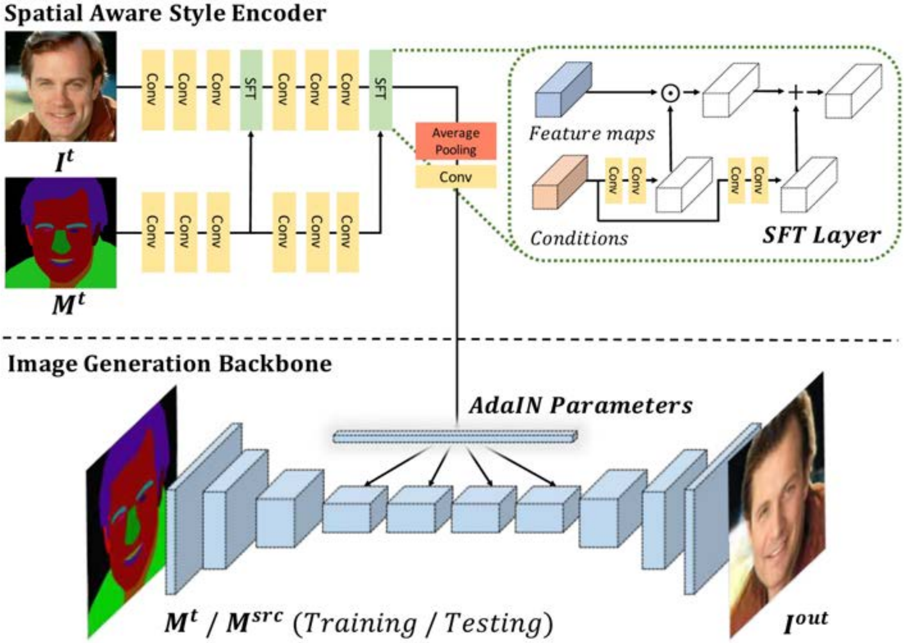
Figure 7 Faceshifter architecture (see online version for colours)



Source: Li et al. (2019)

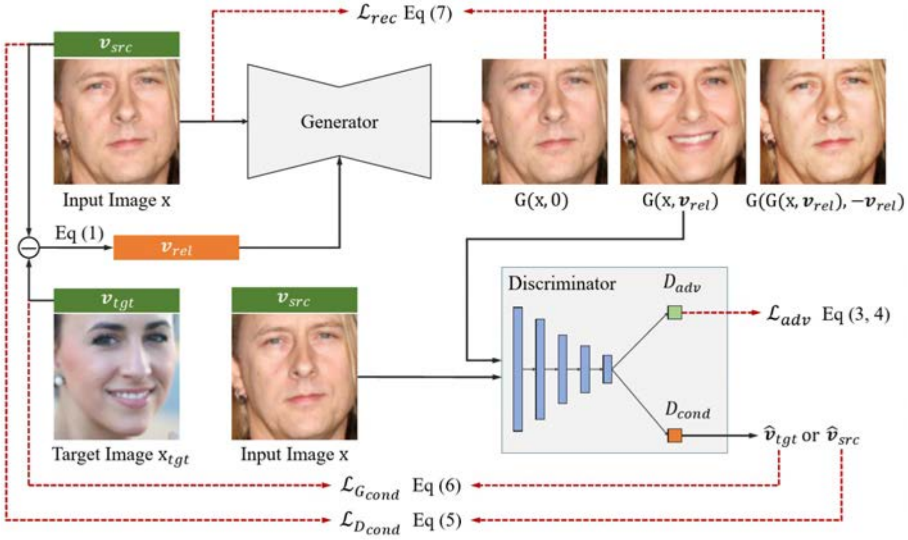
Fidelity is been one of the limitations in the existing facial attribute manipulation methods. To overcome this limitation (Guo et al., 2021) proposed an instance-level facial attribute transfer that not only transfers facial attributes but also transfers style attributes from the source image to the target image using binary attribute annotations. The procedure for accomplishing instance-level face attribute transfer consists of two steps: first, the original characteristics from the target image are removed, then the style attributes taken from the source image are added. A module for untangling styles information from the source image is designed which is given as input to the generator for adding the style attributes in the generation of the new image. The advantage of this approach is that both instance and semantic level attribute editing has been performed thus obtaining promising results. Figure 10 shows the proposed framework.

Figure 8 MaskGAN for facial attribute manipulation (see online version for colours)



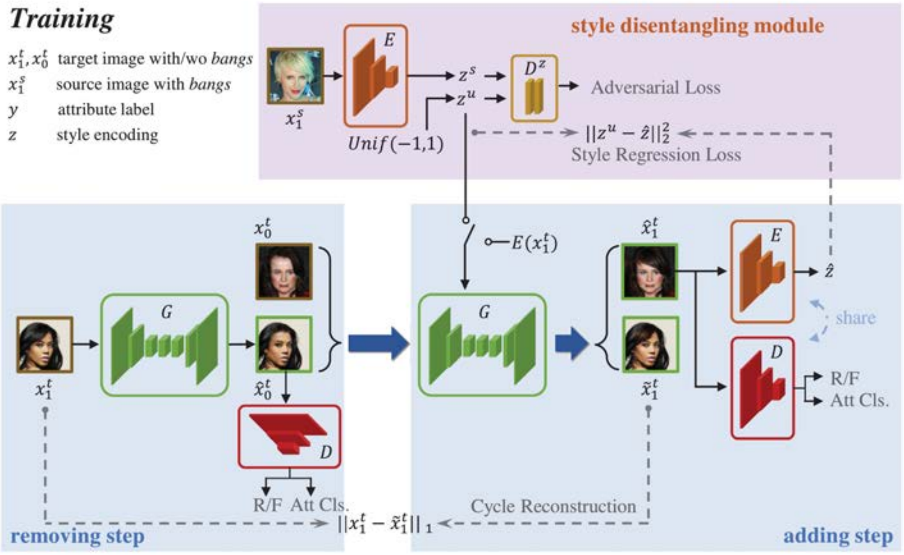
Source: Lee et al. (2020)

Figure 9 Multi-scale fusion approach for facial attribute manipulation (see online version for colours)


$\mathcal{L}_{rec}$ Eq (7)
v_{src}
Input Image x
Generator
G(x, 0)
G(x, v_{rel})
G(G(x, v_{rel}), $-v_{rel}$)
Eq (1)
v_{rel}
v_{tgt}
Target Image x_{tgt}
v_{src}
Input Image x
Discriminator
D_{adv}
D_{cond}
$\mathcal{L}_{adv}$ Eq (3, 4)
$\hat{v}_{tgt}$ or $\hat{v}_{src}$
$\mathcal{L}_{G_{cond}}$ Eq (6)
$\mathcal{L}_{D_{cond}}$ Eq (5)

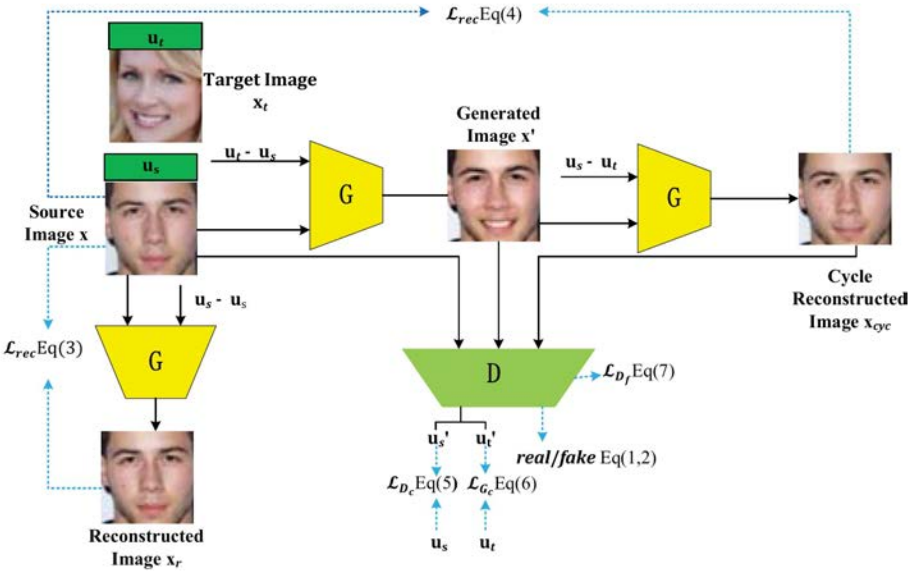
Source: Ling et al. (2020)

Figure 10 STD-GAN framework for facial attribute manipulation (see online version for colours)



Source: Guo et al. (2021)

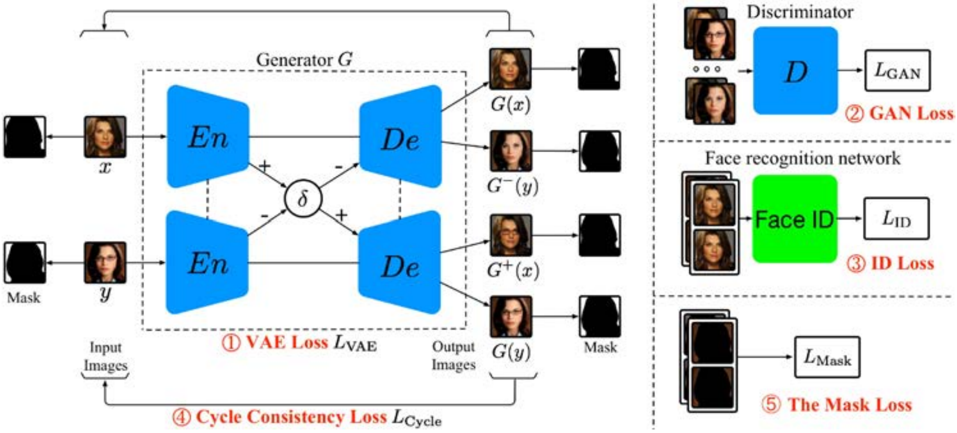
Figure 11 Attention-based GAN for facial attribute manipulation (see online version for colours)



Source: Wang et al. (2021)

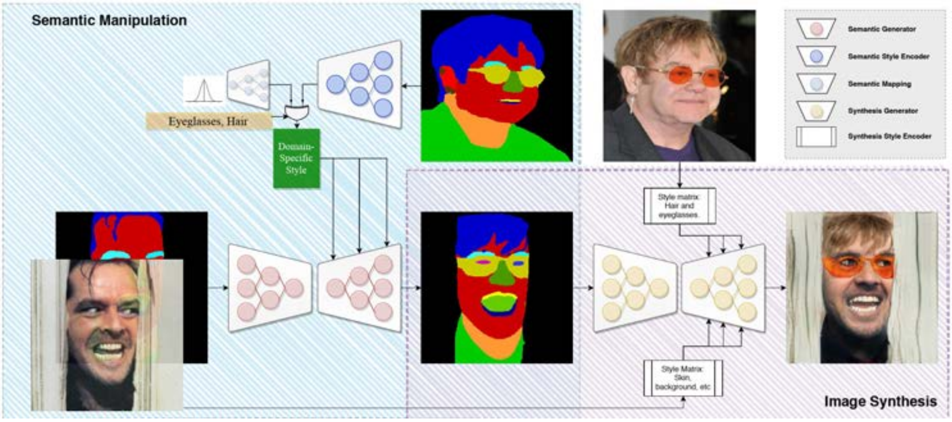
For facial expression manipulation (Wang et al., 2021) developed a conditional GAN model which employed a multi-level attention mechanism that helped in expression manipulation and identity preservation as shown in Figure 11. For generating images with higher quality, feature-based loss and self-attention mechanisms have been used.

Figure 12 Mask adversarial autoencoder approach for facial attribute manipulation (see online version for colours)



Source: Sun et al. (2021)

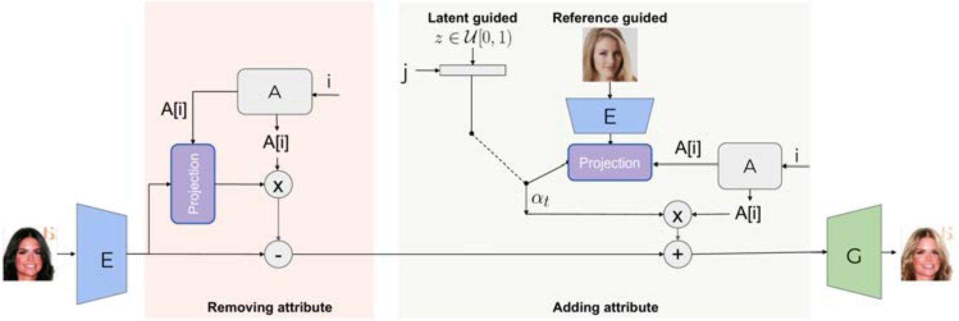
Figure 13 SMILE architecture for facial attribute manipulation (see online version for colours)



Source: Romero et al. (2021)

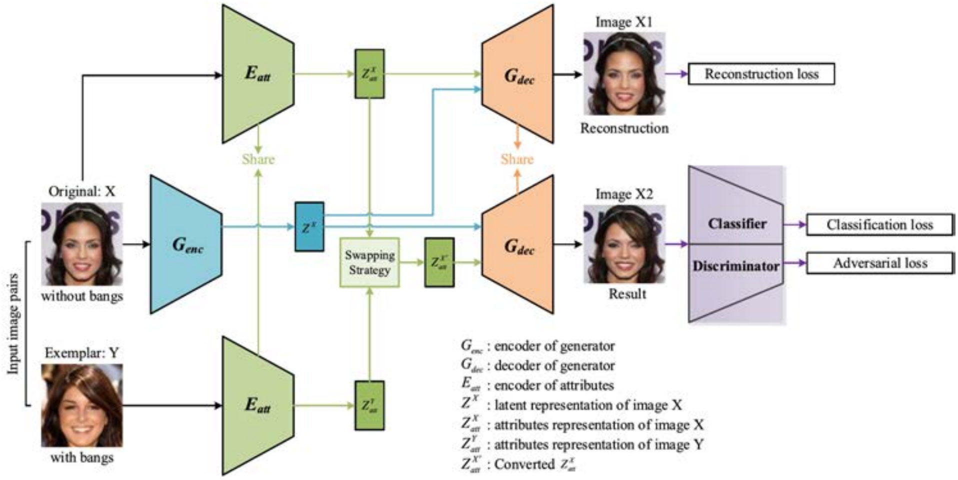
For editing facial attributes, Sun et al. (2021) presented a mask adversarial autoencoder approach that extends the VAE-GAN framework where a few pixels in an encoder's feature maps were altered without affecting the overall information. The authors introduced cycle consistency and facial recognition loss for preserving face details and mask loss to maintain background consistency. Using this approach, they could generate realistic images with varying attributes. The architecture of the proposed model is shown in Figure 12.

Figure 14 Architecture of VecGAN (see online version for colours)



Source: Dalva et al. (2022)

Figure 15 Pluralistic facial attribute editing (see online version for colours)

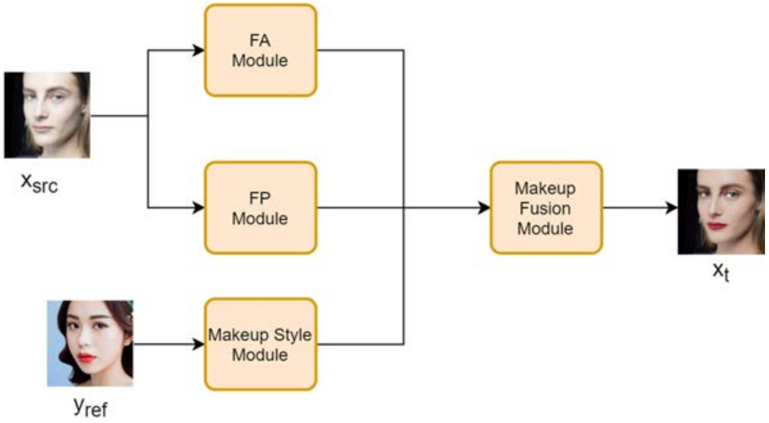


Source: Sun et al. (2022)

Using an image guiding reference approach, Romero et al. (2021) proposed a multi-attribute image-to-image translation using semantic segmentation. Figure 13 depicts the framework of the proposed method, which is an extension of StyleGAN2 (Karras et al., 2019), which deals with semantic masks for performing exemplar-based synthesis.

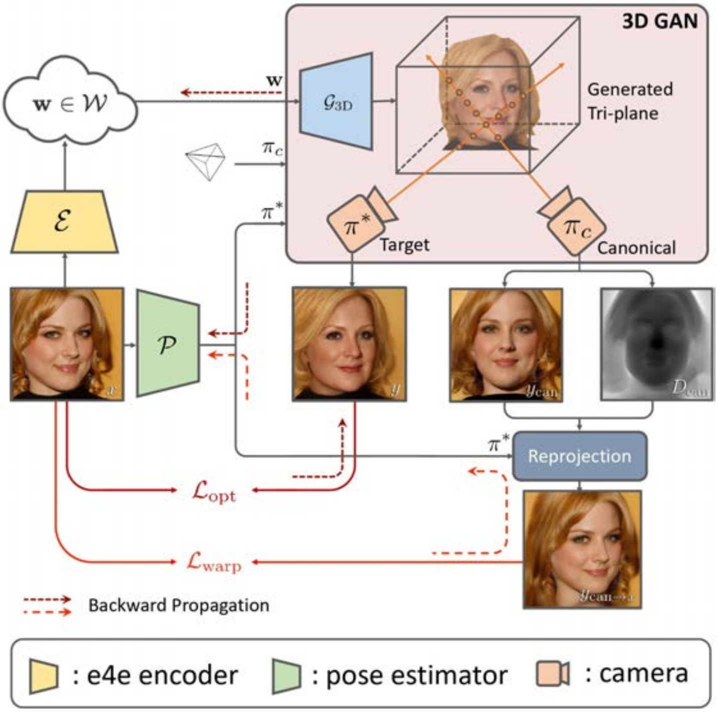
With interpretable latent directions, Dalva et al. (2022) provides VecGAN as shown in Figure 14, an image-to-image translation system for modifying facial attributes. By factoring latent space, we create the attribute editing and discover an orthogonal linear direction for each attribute. The additional factor is the change's scalar, modifiable strength. The other component is the controlled strength of the change, which is a scalar value that can be sampled or encoded from a reference image through projection. VecGAN has been fully trained for image translation tasks and is capable of changing one characteristic while retaining the others. As opposed to earlier works, it uses a single deep encoder-decoder architecture to translate data instead of a separate style network.

Figure 16 Architecture of large-pose facial makeup transfer based on GAN (see online version for colours)



Source: Chang et al. (2018)

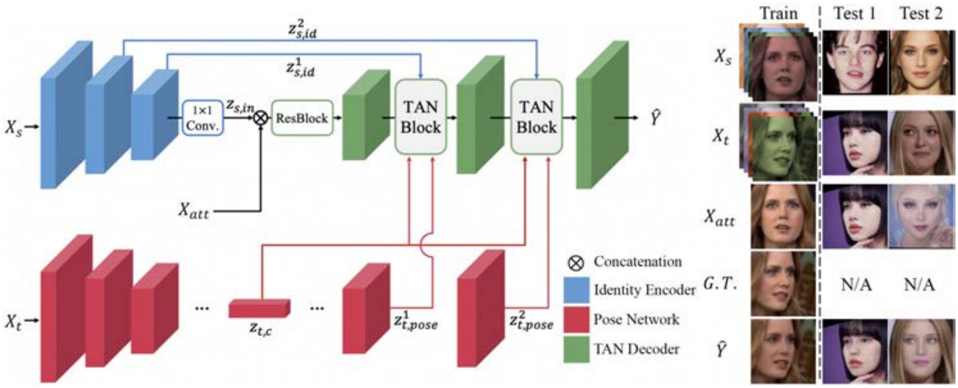
Figure 17 Architecture of 3D GAN inversion with pose optimisation (see online version for colours)



Source: Ko et al. (2023)

The diversity of facial attribute editing is mostly ignored by present approaches, which can only produce a single editing result via binary attribute labels and cannot disclose the diversity of attribute styles. Instead, they concentrate primarily on enhancing the quality of facial attribute editing. To address this restriction, Sun et al. (2022) offers PattGAN as depicted in Figure 15, a unique pluralistic facial attribute editing technique (pluralistic attribute GAN). Rather than using binary attribute labels directly to guide facial attribute editing, PattGAN first learns the disentangled representation of face attributes from binary attribute labels and then utilises the disentangled representation of attributes to assist facial attribute editing. To extract distributions of specific properties from face photos, an independent encoder known as the 'attribute encoder' is introduced. Additionally, by improving the model's capacity to learn pluralistic attributes, a unique swapping technique is created to help the attribute encoder in modelling the disentangled representation of facial characteristics. The attribute encoder, when used with the classification loss, can accurately separate attribute-related information from face photos.

Figure 18 Architecture of FastSwap (see online version for colours)

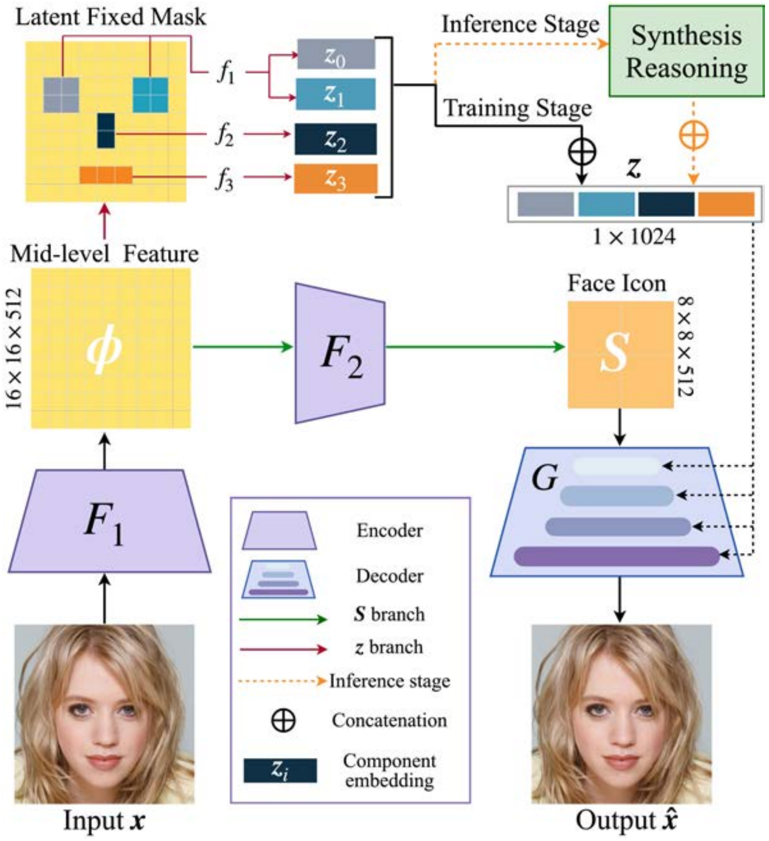


Source: Yoo et al. (2023)

Li and Tu (2023) presented a large-pose makeup transfer approach based on GANs by adopting CycleGAN (Chang et al., 2018) as shown in Figure 16. To locate the crucial areas, including the eyes, mouth, and skin, a face alignment module (FAM) is first introduced. The raw image is then examined to extract the facial features using a face parsing module (FPM) and face parsing losses. The makeup transfer is then finished by fusing the face characteristics and makeup style codes that were derived from the reference image. Cycle consistency loss, perceptual loss, adversarial loss, face parsing loss, and makeup loss are the several loss functions evaluated for evaluation. Although the effect of makeup transfer is improved by this method, it still has the drawback of being unable to transfer the pattern portion of the face makeup.

Ko et al. (2023) provides a 3D-GAN inversion technique that incrementally improves an image's latent code and 3D camera posture as depicted in Figure 17. We extend the recently described 2D GAN inversion approach, which first inverts the provided image into a pivot code and then slightly tunes the generator based on the fixed pivot code [i.e., pivotal tuning (Roich et al., 2022)], resulting in significant results in both reconstruction and editability. Recognising the interdependency between the latent code and camera parameter, we employ a hybrid learning and optimisation-based technique by using an encoder to derive a preliminary approximation of the camera posture and latent code before further refining it to an ideal outcome. Also, the authors use regularisation loss, which makes use of conventional depth-based image warping, to further enforce the camera viewpoint's proximity (Zhou et al., 2017b).

Figure 19 Architecture of IA-FaceS (see online version for colours)

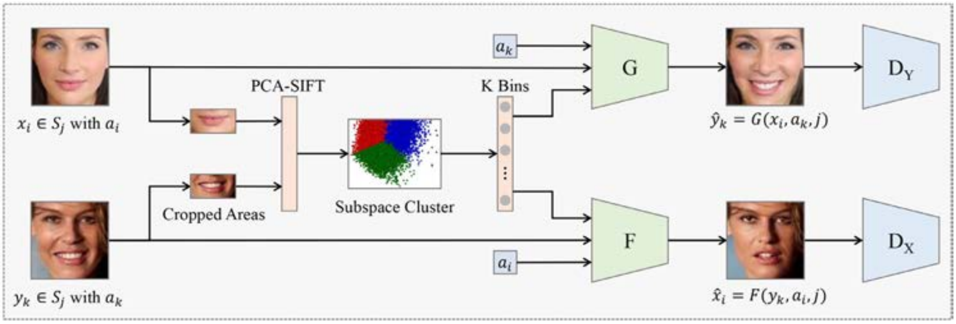


Source: Huang et al. (2023)

Using a TAN block to maintain the identity, pose, and attributes of the inputs as shown in Figure 18, Yoo et al. (2023) suggested a simple one-stage system called FastSwap. The authors use adaptive normalising to tackle the low-fidelity issue that arises as a result of network reduction. To extract the attributes from the target image, a unique data augmentation and switch-test technique are proposed, allowing for controlled attribute manipulation.

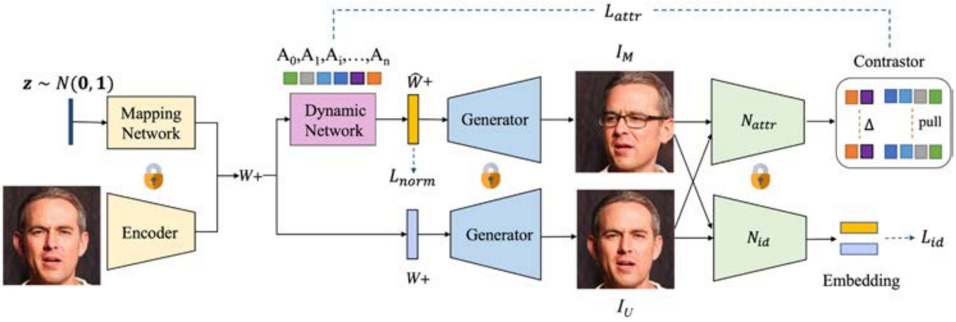
Huang et al. (2023) presented a bidirectional approach for disentangled face attribute modification in addition to flexible, controllable component editing. As shown in Figure 19, images are embedded onto two branches: one computes high-dimensional component-invariant content embedding for capturing facial information, and the other offers low-dimensional component-specific embeddings allowing component manipulations. The two-branch technique allows for high-quality facial component-level editing while maintaining faithful reconstruction of details. Additionally, the component adaptive modulation (CAM) module was developed that successfully separates highly correlated face components by integrating component-specific guidance into the decoder. Without using face masks or sketches, single-eye editing is created for the first time.

Figure 20 Architecture of SC-GAN (see online version for colours)



Source: Li et al. (2023b)

Figure 21 Architecture of DyStyle (see online version for colours)



Source: Li et al. (2023a)

Li et al. (2023b) create a GAN based on subspace clustering (SC-GAN) as shown in Figure 20. Our SC-GAN can simultaneously divide various subspaces and generate different samples, allowing the training of generative models to be more effectively directed by facial attributes and their breakdown and modification in a realistic and significant way. It makes use of the SIFT K-means cluster, which may divide the overall semantic facial space into several subspaces without supervision and aid the new GAN in producing more convincing results in particular subspaces.

To execute nonlinear and adaptive manipulation of latent codes for flexible and precise attribute control, Li et al. (2023a) propose a dynamic style manipulation network (DyStyle) whose structure and parameters alter depending on input samples as shown in Figure 21. The authors suggest the dynamic multi-attribute contrastive learning (DmaCL) approach, which simultaneously decouples several attributes from the generative picture and latent space of the model. This approach will enable efficient and stable optimisation of the DyStyle network. As a result, it displays fine-grained disentangled modifications across a variety of numeric and binary properties. Comparisons with existing style modification approaches, both qualitative and quantitative, demonstrate advantages in terms of multi-attribute control accuracy and identity retention without sacrificing photorealism.

3.2 Facial attribute manipulation approaches using heterogeneous images

Kazemi et al. (2018) presented a conditional CycleGAN (cCycleGAN) as shown in Figure 22 for facial attribute modification for face-sketch synthesis problem. The training was performed using two datasets which included hand-drawn and synthetic sketches in an unpaired manner. For every single sketch, the proposed method generated multiple photos with different facial attributes. The main drawback of this approach was that during the generation, additional structural modifications were observed in certain areas.

Table 4 Overview of heterogeneous facial attribute manipulation approaches

Methods	*Description*	*Datasets*	*Limitations*
Conditional CycleGAN (Kazemi et al., 2018)	A novel framework for facial attribute guided Sketch-Photo synthesis by adding conditions to the CycleGAN.	FERET (Phillips et al., 2000) and CelebA (Liu et al., 2015)	During the generation, additional structural modifications were observed in certain areas.
Attribute-guided generative discriminative network (Shiri et al., 2019)	Encodes stylised images with facial attributes and then recovers realistic faces from encoded feature maps using an autoencoder with residual block-embedded skip-connections to extract residual feature maps.	CelebA (Liu et al., 2015)	Pre-defined facial attributes are considered.
Encoder-decoder architecture (Hu and Guo, 2020)	A convolutional generative network with encoder-decoder architecture is proposed to achieve the facial attribute-controlled face sketch to image translation.	CelebA (Liu et al., 2015)	Pre-defined facial attributes are considered.
Joint sketch attribute learning approach (Yang et al., 2020a)	A novel approach that synthesises photo-realistic face images from sketches and attributes using joint sketch-attribute learning.	CelebA-HQ (Karras et al., 2017)	When huge distortions in the image exist, generated images were found to be of low quality.
Sketch-to-image (Phusomsai and Limpiyakorn, 2020)	A face sketch to image translation by manipulating a single facial attribute, wavy hair, straight hair, wearing glasses.	CelebA (Liu et al., 2015)	Pre-defined facial attributes are considered.

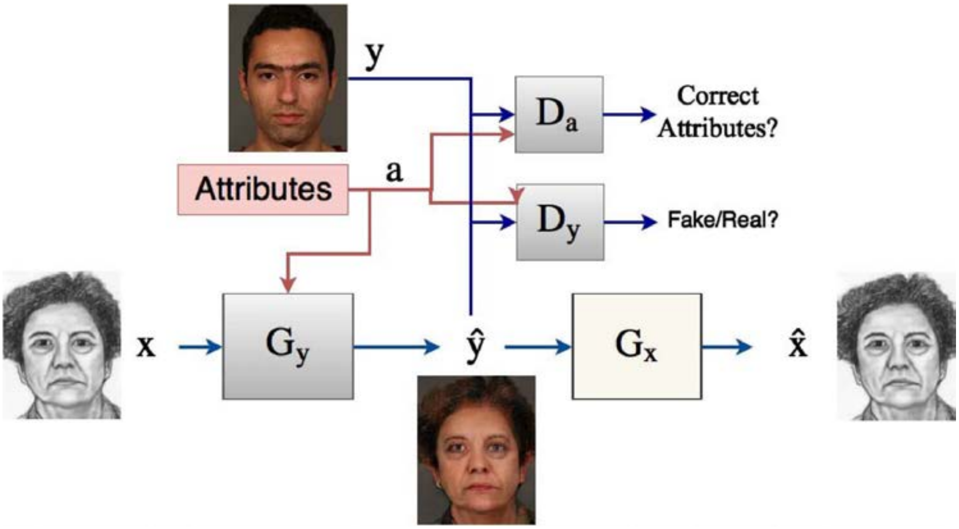
Table 4 Overview of heterogeneous facial attribute manipulation approaches (continued)

Methods	*Description*	*Datasets*	*Limitations*
S2FGAN (Yang et al., 2020b)	For sketch-to-image translation with the facility of face reconstruction, attribute editing, and interactive manipulation of attribute intensity and present a semantic level perceptual loss to increase the sketch-to-image translation quality.	CelebAMask-HQ (Karras et al., 2017)	Pre-defined semantic masks are considered.
r-face (Deng et al., 2020)	Reference guided face component editing for diverse and controllable face component editing with geometric changes using an example-guided attention module which breaks the shape and intermediate presentation.	CelebAMask-HQ (Karras et al., 2017)	Difficulty with reference images with considerable changes in pose.
Controllable sketch-to-image (Yang et al., 2021)	A style-based network architecture for sketch-to-image translation which adapts edge-based models to real-world hand-drawn sketches using the sketch refinement method.	CelebA-HQ (Karras et al., 2017) and CelebA (Liu et al., 2015)	Synthesised images can be generated by repeated revision until satisfied.
CMAFGAN (Luo et al., 2022)	Cross-modal attention fusion-based generative adversarial network for attribute word-to-face synthesis.	CelebA (Liu et al., 2015) and LFW (Huang et al., 2007)	Some irregular and fuzzy images exist in synthesised images which require more improvement.
A bi-directional facial attribute transfer framework (Shi et al., 2022)	A bi-directional facial attribute transfer framework for transferring facial attribute to portrait illustration.	CelebA (Liu et al., 2015)	Only single attribute can be changed at a time.
FLAME (Parihar et al., 2022)	Attribute editing and attribute style manipulation by StyleGAN latent space exploration.	CelebAMask-HQ (Karras et al., 2017)	Generating synthetic image pairs can be challenging and can be misused.
Pastiche Master (Yang et al., 2022)	An exemplar-based high-resolution portrait style transfer using DualStyleGAN.	CelebA-HQ (Karras et al., 2017)	Pre-defined semantic masks are considered.
SM-GAN (Chen et al., 2022)	A generative adversarial network (GAN) with semantic masks.	CelebA (Liu et al., 2015) and LFW (Huang et al., 2007)	Pre-defined semantic masks are considered.

Table 4 Overview of heterogeneous facial attribute manipulation approaches (continued)

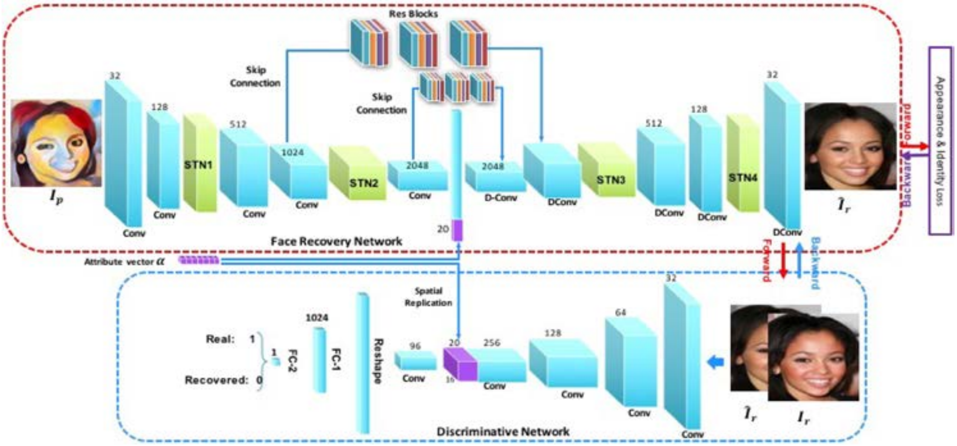
Methods	*Description*	*Datasets*	*Limitations*
Sketch2Photo (Liu et al., 2023)	A novel image generation method that can create photorealistic images from weak or incomplete sketches or edge maps while simultaneously capturing global contexts and local details.	CelebA-HQ (Karras et al., 2017)	The quality of the image generated may degrade if the alignment of images is too large.
HIGSA (Wu et al., 2023)	The human image generation framework consists of self-attention blocks referred to as the stripe self-attention block (SSAB) and the content attention block (CAB) to produce photo-realistic human images.	Market-1501 (Zhu et al., 2019b)	Only 18 human key points were extracted by the human pose estimator. The number of training parameters is more which forms a complex neural network.
3D avatar generation (Canfes et al., 2023)	A novel 3D modification technique that makes use of the contrastive language-image pre-training (CLIP) model and a pre-trained 3D GAN model to produce face avatars.	-	Applicable only to TBGAN.

Figure 22 cCycleGAN for facial attribute modification (see online version for colours)



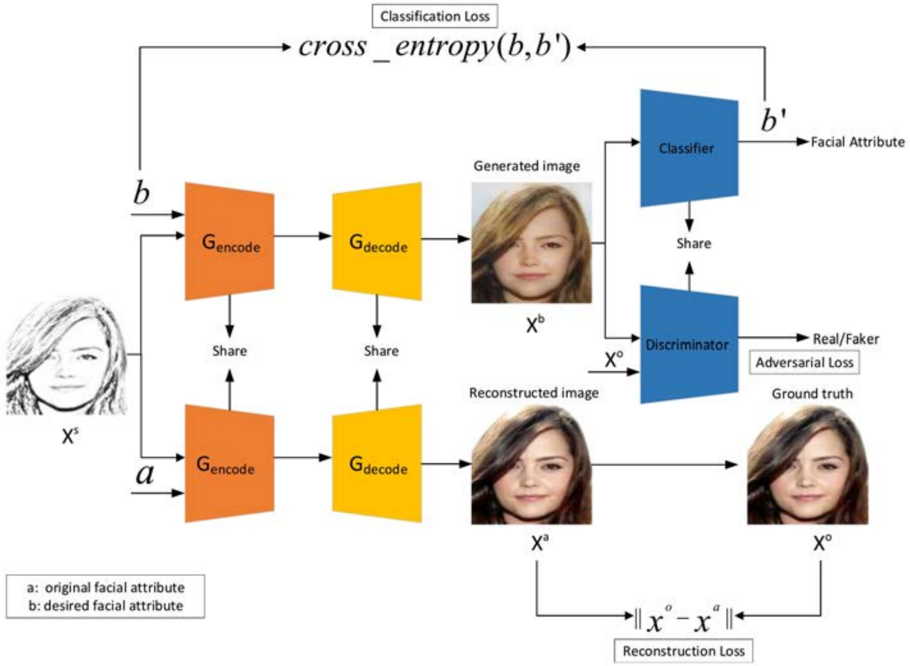
Source: Kazemi et al. (2018)

Figure 23 Recovering faces from portraits (see online version for colours)



Source: Shiri et al. (2019)

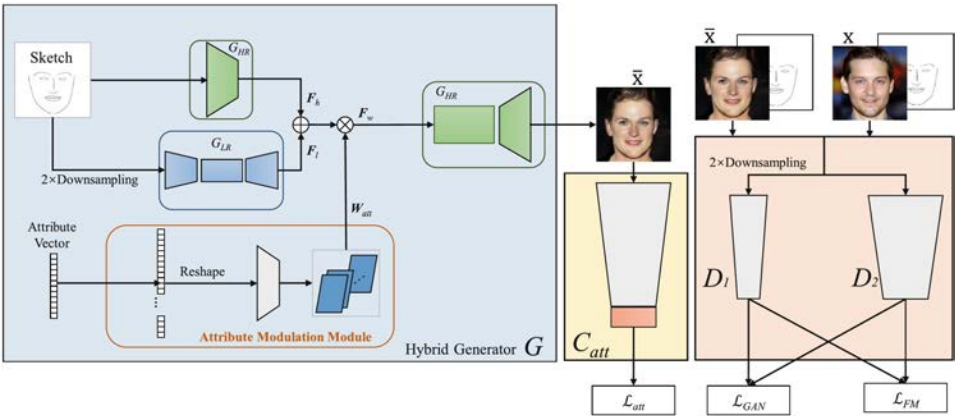
Figure 24 Encoder-decoder architecture for the generation of facial images from sketches (see online version for colours)



Source: Hu and Guo (2020)

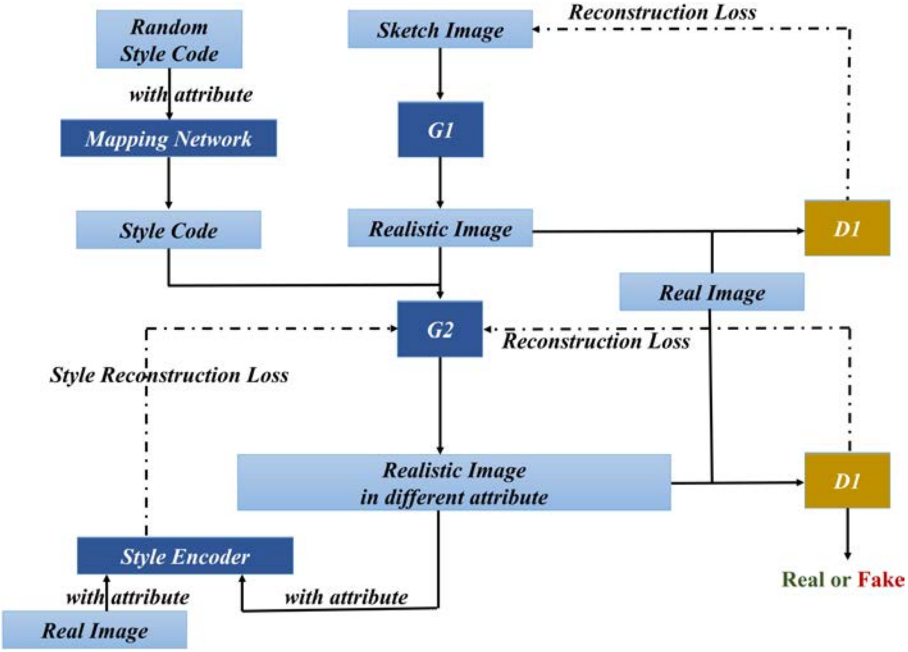
Shiri et al. (2019) proposed an attribute-guided generative-discriminative network approach as depicted in Figure 23 to recover realistic photos from unaligned portraits, real paintings, and hand-drawn sketches. Autoencoder with residual block-embedded skip-connections was used for extracting residual features from portraits (visible features) and was combined with the facial attributes (semantic information). Thus manipulating the attribute vectors, realistic faces were generated with desired facial attributes.

Figure 25 Joint sketch attribute learning approach for the generation of facial images (see online version for colours)



Source: Yang et al. (2020a)

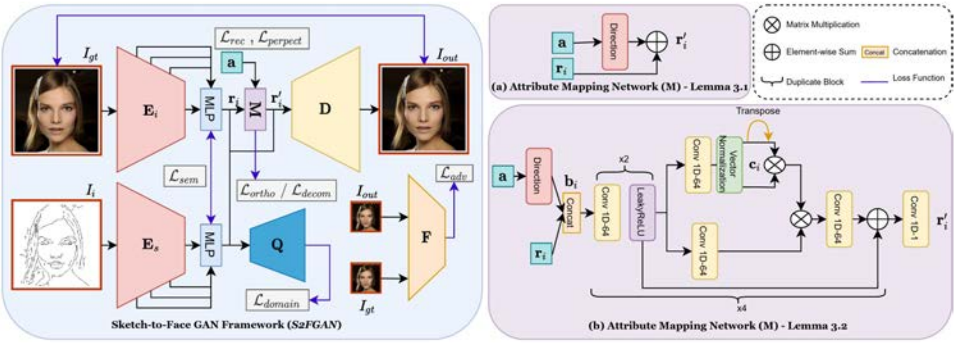
Figure 26 Generating images from sketches based on GAN's (see online version for colours)



Source: Phusomsai and Limpiyakorn (2020)

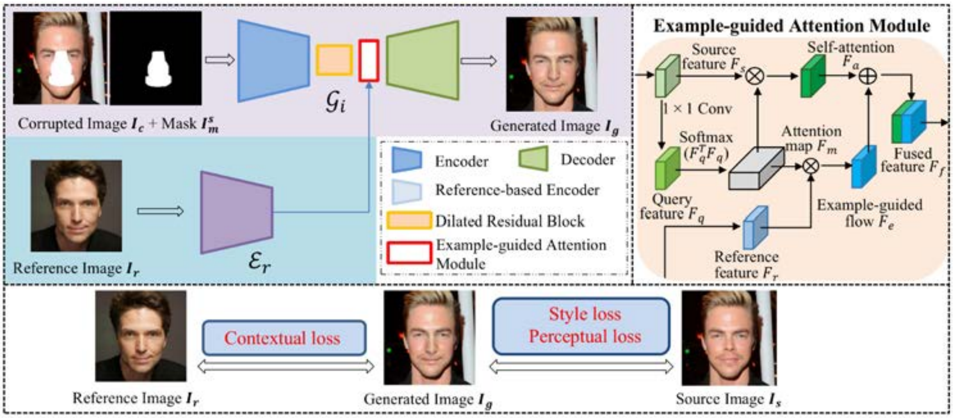
Hu and Guo (2020) proposed an encoder-decoder architecture for the generation of facial images from sketches as shown in Figure 24. This architecture was able to control multiple attributes without affecting other facial attributes and generate high-quality images. The architecture design was similar to AttGAN (He et al., 2019) with the addition of more residual blocks for extracting the facial details of the generated images. Attribute classification loss to guarantee that the reconstructed face image has the facial characteristics that the users want, reconstruction loss to combine information about the texture and structure of the face, and adversarial loss to promote visual fidelity, together constituted in obtaining high-quality images from sketches.

Figure 27 Network architecture of S2FGAN framework (see online version for colours)



Source: Yang et al. (2020b)

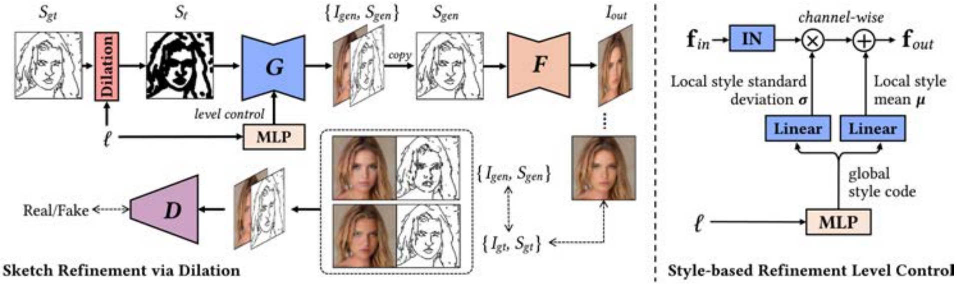
Figure 28 r-face for facial attribute manipulation (see online version for colours)



Source: Deng et al. (2020)

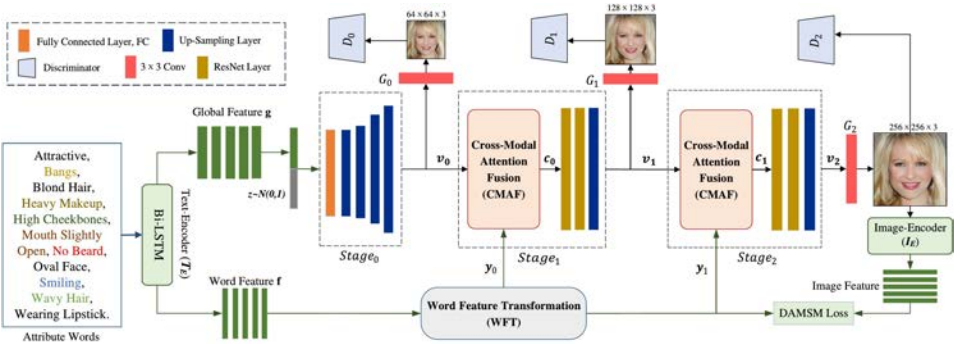
The joint sketch attribute learning approach was designed by Yang et al. (2020a) using conditional GANs for generating face images from the embedding of geometric shapes. These shapes provided details of the facial structure and details about the facial attributes. An attribute modulation module that transfers user-preferred attributes to augment sketch representation with details was suggested and were able to create face images with finer control over both the shape and appearance of the face. One of the limitations of this approach is that when huge distortions in the image exist generated images were found to be of low quality. The architecture of the proposed model is shown in Figure 25.

Figure 29 Controllable sketch-to-image translation for facial attribute manipulation (see online version for colours)



Source: Yang et al. (2021)

Figure 30 Architecture of CMAF-based GAN for attribute word-to-face synthesis (see online version for colours)

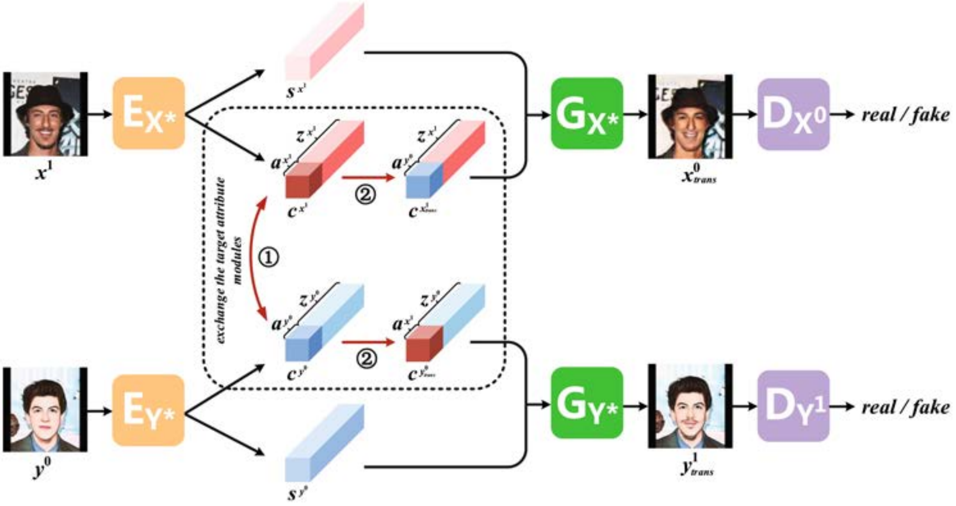


Source: Luo et al. (2022)

Phusomsai and Limpiyakorn (2020) suggested a method of generating images from sketches based on GANs such as Pix2Pix (Isola et al., 2017) and StarGan2 (Choi et al., 2020). Pix2Pix (Isola et al., 2017) is considered an effective method for synthesising, reconstructing, and coloring black and white images, and StarGan2 (Choi et al., 2020) helps in mapping the facial attributes along with the style information between source and generated images. The model was able to generate realistic images with single attribute manipulation such as straight hair, wavy hair, and images with glasses. The architecture is depicted in Figure 26.

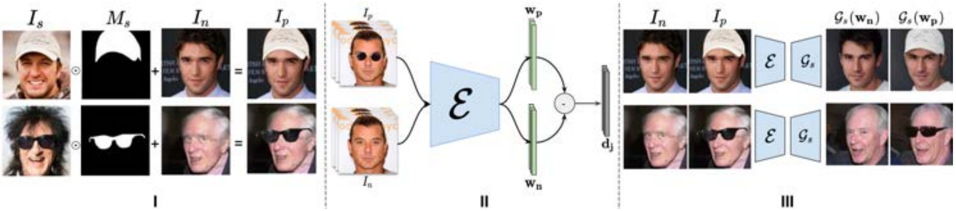
An interactive facial image manipulation approach using sketches was developed by Yang et al. (2020b) where attribute manipulations can be controlled on either sketches or facial masks. As shown in the Figure 27 for the generation of faces from sketches, encoder-decoder GAN architecture has been utilised. As the sketches contain less information identifying low-level facial features is a challenging task. Hence, the semantic level perceptual loss has been proposed in identifying the low-level features of the sketches and also the structure of the face. Two attribute mapping networks with latent semantic loss were used to adjust semantics in the latent space for attribute editing which preserves the semantics and strength of non-edited attributes.

Figure 31 A bi-directional facial attribute transfer framework for transferring facial attribute to portrait illustration (see online version for colours)



Source: Shi et al. (2022)

Figure 32 Attribute editing and attribute style manipulation by StyleGAN latent space exploration (see online version for colours)

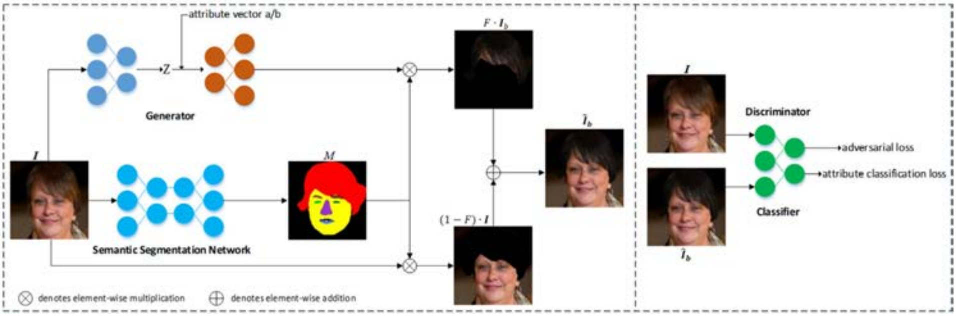


Source: Parihar et al. (2022)

To overcome the limitations of the existing methods where pre-defined facial attributes were considered for modification or considering manually edited masks or sketches, Deng et al. (2020) proposed a framework called r-face for modifying the facial attributes. As a backbone for this model, an image in-painting model is used, with reference images serving as conditions for controlling the shape of face components.

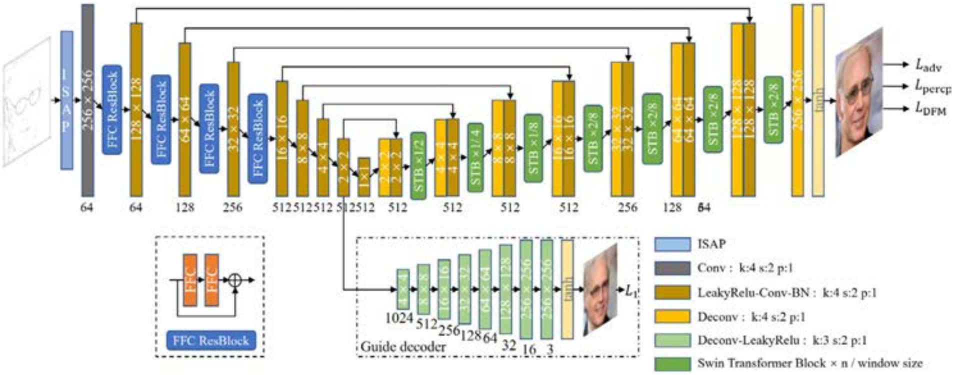
Yang et al. (2021) proposed a style-based network architecture as shown in Figure 29 that learns to refine the sketches. The sketch refinement process uses coarse-to-fine dilation inspired by the painting process of artists which bridges the gap between coarse-level sketches and fine-level edges.

Figure 33 Harnessing semantic segmentation masks for accurate facial attribute editing (see online version for colours)



Source: Chen et al. (2022)

Figure 34 Synthesising photo-realistic images from sketches via global contexts (see online version for colours)

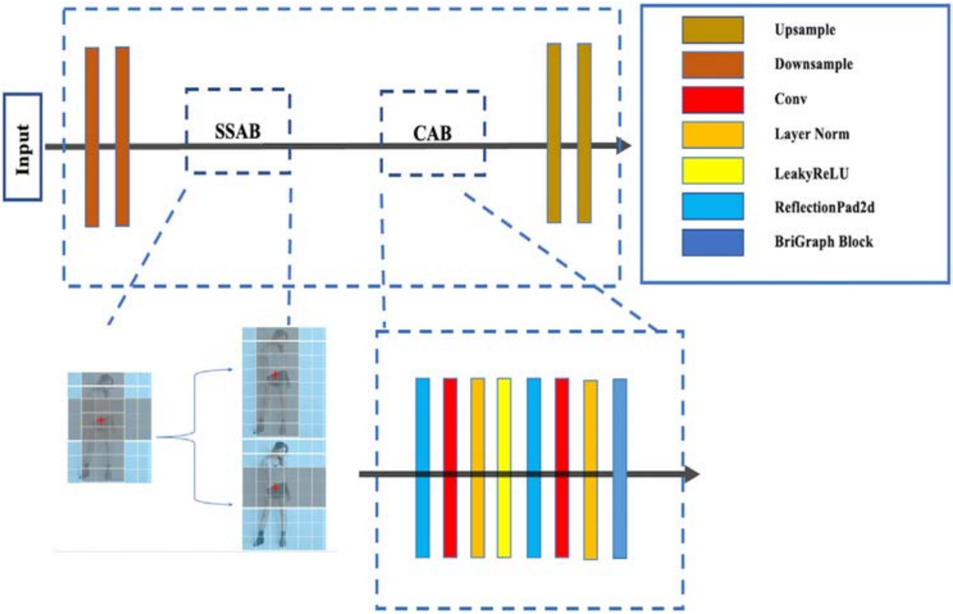


Source: Liu et al. (2023)

Instead of using the real attribute vectors, CMAFGAN (Luo et al., 2022) creates faces using the matching attribute words where the word feature transformation (W2F) challenge has never been satisfactorily completed before as shown in Figure 30. The two blocks of CMAFGAN that are recommended to examine the association between image data and the equivalent attribute word features are cross-modal attention fusion (CMAF) and word feature transformation (WFT). By utilising cross-attention, CMAF provides a more effective strategy than past studies for fusing word and visual characteristics.

In Shi et al. (2022), the newer and more difficult challenge of facial attribute transfer between diverse photos is tackled. For instance-based facial attribute transfer, the authors describe a new bi-directional method based on GAN and latent representation that intends to transfer a target facial attribute with its fundamental shape from a reference photo-realistic face image to a source realistic portrait drawing and vice versa as shown in Figure 31. The method solves the new facial attribute transfer task by redefining and reformulating an image's latent representation, which combines an image's content and aesthetic style in a representation. In a supervised manner, it disentangles each supplied image into two components to generate associated content latent representation and visual style representation. Not only is the facial attribute transferred from a photo-realistic facial image to a realistic portrait illustration accomplished by exchanging the target attribute modules in content latent representations of two heterogeneous images, but the target attribute is also removed from the reference photo-realistic facial image.

Figure 35 Human image generation with self-attention (see online version for colours)

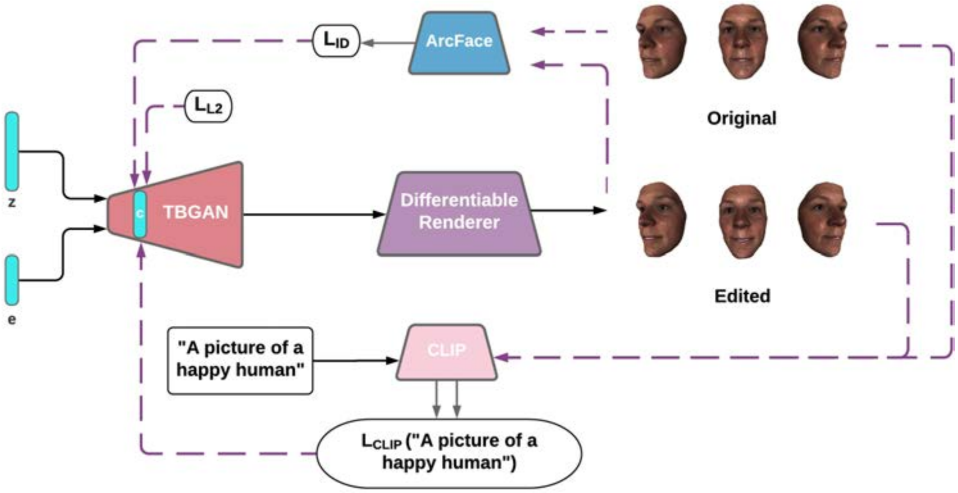


Source: Wu et al. (2023)

Few-shot latent-based attribute manipulation and alteration (FLAME) as shown in Figure 32 is a simple yet effective framework for doing highly controlled picture editing via latent space manipulation, according to Parihar et al. (2022). To manipulate semantic features in the resulting image, they estimate linear directions in the latent space (of a pre-trained StyleGAN). They offer a novel task of attribute style manipulation to generate varied styles for characteristics such as eyeglasses and hair, as well as a novel sampling approach to sample latent from the manifold, allowing to generate a diverse collection of attribute styles in addition to those in the training set. One constraint of this technique is the difficulty in curating synthetic image pairs in some circumstances, such as gender modification.

In this work, Yang et al. (2022) investigates more difficult exemplar-based high-resolution portrait style transfer by presenting a novel DualStyleGAN with variable management of dual styles of the original face domain and the extended artistic portrait domain. With just a few hundred style examples, a novel DualStyleGAN is proposed to characterise and manipulate intrinsic and extrinsic styles for exemplar-based high-resolution portrait style transfer. This method outperforms state-of-the-art approaches in the creation of high-quality and varied artistic portraits.

Figure 36 Text and image guided 3D avatar generation and manipulation (see online version for colours)



Source: Canfes et al. (2023)

The three main challenges with modifying face attributes are accurate editing area, invariant identification information, and realistic visual effect. Unfortunately, most studies concentrate on the first two issues. However, the main cause of attribute-irrelevant data being damaged is a lack of understanding of the accurate editing area in the assignment. Chen et al. (2022) offers a unique face attribute editing algorithm – a GAN with semantic masks – to handle this problem from the standpoint of modifying location accuracy as shown in Figure 33. The semantic segmentation network can only confine manipulation in the target zone by creating the mask to attribute-related areas. The complete framework, known as SM-GAN, is created by combining the GAN with the semantic segmentation network.

Liu et al. (2023) presents a Sketch2Photo, a novel image generation method that can create photorealistic images from weak or incomplete sketches or edge maps while simultaneously capturing global contexts and local details as shown in Figure 34. The lowest layers of the network are constructed using fast Fourier convolution (FFC) residual blocks to provide global receptive fields, then swin transformer block (STB) units are added to effectively get long-range global contexts for large-scale feature maps. Additionally, an enhanced spatial attention pooling (ISAP) module is provided to ease the rigorous alignment constraints between generated images and incomplete sketches. Extensive experimental results and comparisons reveal that the suggested method is capable of achieving excellent image synthesis effects and outperforms several other existing methods. During the training procedure, adversarial loss, reconstruction loss, feature matching loss, and perceptual loss are assessed. The quality of the image generated whenever the alignment of the face is too large may result in a poor image, which is one of the drawbacks of the suggested approach. Furthermore, the quality of the created image may decline due to the presence of things in the environment, such as pedestrians surrounded by trees.

By using the positional data from the source image, Wu et al. (2023) suggest a human image generation framework termed human image generation with self-attention (HIGSA) as shown in Figure 35. Two complementary self-attention blocks referred to as the stripe self-attention block (SSAB) and the content attention block (CAB), are each included in the proposed HIGSA to produce photo-realistic human images. The attention map is computed for each pixel in SSAB based on its relative spatial positions to other pixels and establishes global dependencies of human images and introduces a powerful feature extraction module in CAB that can be used to interactively improve both person's look and shape feature representations. Because of this, the HIGSA framework automatically maintains superior shape consistency and visual consistency with finer details. HIGSA can synthesise human images in any posture with realistic details and maintained features, according to both qualitative and quantitative results. During the model evaluation, GAN loss, perceptual loss, and combined L1 loss are all measured. Some of the limitations of this approach are that only 18 human key points are extracted by human pose estimator (Zhu et al., 2019b) which may not result in accurate pose representation and hence may be difficult to extend to other complex computer vision fields. Also, the number of training parameters considered results in a complex neural network.

By employing text- or image-based instructions like 'a young face' or 'a surprised face', Canfes et al. (2023) suggest a novel 3D modification technique that can change the model's texture as well as its shape as shown in Figure 36. To produce face avatars and build a completely differentiable rendering pipeline, they make use of the contrastive language-image pre-training (CLIP) model and a pre-trained 3D GAN model. More precisely, this takes an input latent code and adjusts it so that the target attribute indicated by a text or image prompt is present or improved while mostly preserving the integrity of other properties. Using this method, editing an image just takes five minutes. However, this method has certain drawbacks, including the fact that the model has only been trained to build partial facial avatars and that full-head mesh creation is not possible. Also, this approach can be implemented using only TBGAN (Gecer et al., 2020).

4 Challenges and opportunities

Even though promising results have been obtained in manipulating facial attributes, still there exist many challenging issues that need to be addressed. Many opportunities are open to the research domain if these challenges are taken into consideration. This section discusses the challenges and opportunities involved in facial attribute manipulation keeping in mind the different databases, models/algorithms, and applications.

During the discussion on deep learning, the first thought considered will be the availability of data. To extract the various facial features, we need to train the model using a maximum number of samples. When a sufficient amount of data is not available, the performance of the deep learning model degrades. By considering the well-known database, CelebA dataset, we can observe that there is an imbalance in the distribution of data under different attribute categories with a minimum of 4,547 images under the 'bald' category of images and a maximum of 169,158 images under 'no beard' category of images. Also, it is observed that only a few datasets with attribute information are available, and only a selected few attributes are considered for facial manipulations. However, only a few attributes have resulted in promising results (Chen et al., 2016; Xiao et al., 2018; Huai-Yu et al., 2018).

The other possible challenge is the availability of video data for facial attribute manipulation. To date, no study exists for the manipulation of facial attributes in videos. Obtaining faces and identifying the attributes of faces to be changed in a video frame and preserving the identity is not an easy task. The other factor that should also be considered is the quality of the videos which can affect the performance.

Yet another challenge is obtaining appealing performance from low-resolution images. Editing facial attributes in low-resolution images is difficult. Currently, many deep learning models use a specific range of resolutions under defined conditions to accurately modify facial attributes, which produces satisfactory results. Thus by using super-resolution techniques on the regenerated images, the performance of the deep learning model can be improved.

Adversarial images, which are created by adding minor artificial changes to the topology of the network, training phase, and hyperparameter variations, can indeed be leveraged as inputs to the deep face attribute prediction model. The strength of models could be enhanced by accurately classifying the original inputs and misclassifying the adversarial inputs. Szegedy et al. (2017) first postulated that properly designed modifications that are invisible to humans can cause neural networks to misclassify an image. Following this discovery, academics are beginning to consider the study of adversarial images.

Additionally, some severe facial attribute manipulation may have a negative impact on an automated face recognition system's performance which is also one of the challenging tasks. Bias caused by naturally occurring demographic characteristics in automated face recognition systems might be exacerbated when attributes are digitally manipulated. For instance, if we attempt to alter sex cues by introducing goatee to images of women and makeup to give men a more feminine appearance and results in visible aberrations that may be responsible for a significant decrease in automated face recognition performance. In yet another attribute change, by removing eyeglasses, it is possible to lighten the lens shades or the color of the eyeglass frames, but not completely remove the glasses. This may result in significant changes in the texture surrounding the facial landmarks, lowering the accuracy of face recognition.

5 Conclusions

In the realm of computer vision, facial qualities play a critical role in recognising the visible features of face images. The performance of many authentic applications has been enhanced using these facial features. In this paper, we discuss an ample review of deep learning models based on facial attribute manipulation, different databases, and metrics used for evaluation. Also, details of various state-of-the-art techniques along with pros, cons, future problems, and opportunities are emphasised. From these investigations, it can be seen that the availability of low-resolution photos and the quality of the images produced are the key limitations of face attribute modification. We look forward to more research that addresses these issues and takes advantage of these opportunities to advance the field of deep-face attribute manipulation.

References

Abdal, R., Zhu, P., Mitra, N.J. and Wonka, P. (2021) 'Styleflow: attribute-conditioned exploration of StyleGAN-generated images using conditional continuous normalizing flows', *ACM Transactions on Graphics (ToG)*, Vol. 40, No. 3, pp.1–21.

Buolamwini, J. and Gebru, T. (2018) 'Gender shades: intersectional accuracy disparities in commercial gender classification', in *Conference on Fairness, Accountability and Transparency*, PMLR, pp.77–91.

Canfes, Z., Atasoy, M.F., Dirik, A. and Yanardag, P. (2023) 'Text and image guided 3D avatar generation and manipulation', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.4421–4431.

Cao, C., Weng, Y., Zhou, S., Tong, Y. and Zhou, K. (2013) 'Facewarehouse: a 3D facial expression database for visual computing', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 20, No. 3, pp.413–425.

Chang, H., Lu, J., Yu, F. and Finkelstein, A. (2018) 'PairedCycleGAN: asymmetric style transfer for applying and removing makeup', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.40–48.

Chen, B-C., Chen, C-S. and Hsu, W.H. (2014) 'Cross-age reference coding for age-invariant face recognition and retrieval', in *European Conference on Computer Vision*, Springer, pp.768–783.

Chen, P., Xiao, Q., Xu, J., Dong, X., Sun, L., Li, W., Ning, X., Wang, G. and Chen, Z. (2022) 'Harnessing semantic segmentation masks for accurate facial attribute editing', *Concurrency and Computation: Practice and Experience*, Vol. 34, No. 12, p.e5798.

Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I. and Abbeel, P. (2016) 'InfoGAN: interpretable representation learning by information maximizing generative adversarial nets', in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp.2180–2188.

Chen, X., Ni, B., Liu, N., Liu, Z., Jiang, Y., Truong, L. and Tian, Q. (2020) 'COOGAN: a memory-efficient framework for high-resolution facial attribute editing', in *European Conference on Computer Vision*, Springer, pp.670–686.

Choi, Y., Choi, M., Kim, M., Ha, J-W., Kim, S. and Choo, J. (2018) 'StarGAN: unified generative adversarial networks for multi-domain image-to-image translation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.8789–8797.

Choi, Y., Uh, Y., Yoo, J. and Ha, J-W. (2020) 'StarGANv2: diverse image synthesis for multiple domains', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.8188–8197.

Chu, W., Tai, Y., Wang, C., Li, J., Huang, F. and Ji, R. (2020) 'SSCGAN: facial attribute editing via style skip connections', in *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XV 16*, Springer, Glasgow, UK, 23–28 August, pp.414–429.

Chung, J.S., Nagrani, A. and Zisserman, A. (2018) *VoxCeleb2: Deep Speaker Recognition*, arXiv preprint arXiv:1806.05622.

Collins, E., Bala, R., Price, B. and Susstrunk, S. (2020) 'Editing in style: uncovering the local semantics of GANs', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5771–5780.

Dalva, Y., Altindis, S.F. and Dundar, A. (2022). *VecGAN: Image-to-Image Translation with Interpretable Latent Directions*, arXiv preprint arXiv:2207.03411.

Deb, D., Nain, N. and Jain, A.K. (2018) 'Longitudinal study of child face recognition', in *2018 International Conference on Biometrics (ICB)*, IEEE, pp.225–232.

Deng, Q., Cao, J., Liu, Y., Chai, Z., Li, Q. and Sun, Z. (2020) *Reference-Guided Face Component Editing*, arXiv preprint arXiv:2006.02051.

Esser, P., Rombach, R. and Ommer, B. (2020) 'A disentangling invertible interpretation network for explaining latent representations', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.9223–9232.

Fu, Y., Hospedales, T.M., Xiang, T., Gong, S. and Yao, Y. (2014) 'Interestingness prediction by robust learning to rank', in *European Conference on Computer Vision*, Springer, pp.488–503.

Gecer, B., Lattas, A., Ploumpis, S., Deng, J., Papaioannou, A., Moschoglou, S. and Zafeiriou, S. (2020) 'Synthesizing coupled 3D face modalities by trunk-branch generative adversarial networks', in *European Conference on Computer Vision*, Springer, pp.415–433.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) 'Generative adversarial nets', *Advances in Neural Information Processing Systems*, Vol. 27.

Gross, R., Matthews, I., Cohn, J., Kanade, T. and Baker, S. (2010) 'Multi-pie', *Image and Vision Computing*, Vol. 28, No. 5, pp.807–813.

Guan, S., Tai, Y., Ni, B., Zhu, F., Huang, F. and Yang, X. (2020) *Collaborative Learning for Faster StyleGAN Embedding*, arXiv preprint arXiv:2007.01758.

Guo, J., Qian, Z., Zhou, Z. and Liu, Y. (2019) *MulGAN: Facial Attribute Editing by Exemplar*, arXiv preprint arXiv:1912.12396.

Guo, X., Kan, M., He, Z., Song, X. and Shan, S. (2021) 'Image style disentangling for instance-level facial attribute transfer', *Computer Vision and Image Understanding*, Vol. 207, p.103205.

Hand, E., Castillo, C. and Chellappa, R. (2018a) 'Doing the best we can with what we have: multi-label balancing with selective learning for attribute prediction', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1.

Hand, E.M., Castillo, C.D. and Chellappa, R. (2018b) 'Predicting facial attributes in video using temporal coherence and motion-attention', in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp.84–92.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.

He, Z., Zuo, W., Kan, M., Shan, S. and Chen, X. (2019) 'AttGAN: facial attribute editing by only changing what you want', *IEEE Transactions on Image Processing*, Vol. 28, No. 11, pp.5464–5478.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S. (2017) 'GANs trained by a two time-scale update rule converge to a local Nash equilibrium', *Advances in Neural Information Processing Systems*, Vol. 30.

Hu, M. and Guo, J. (2020) 'Facial attribute-controlled sketch-to-image translation with generative adversarial networks', *EURASIP Journal on Image and Video Processing*, No. 1, pp.1–13.

Huai-Yu, L., Wei-Ming, D. and Hu, B-G. (2018) 'Facial image attributes transformation via conditional recycle generative adversarial networks', *Journal of Computer Science and Technology*, Vol. 33, No. 3, pp.511–521.

Huang, G.B., Ramesh, M., Berg, T. and Learned-Miller, E. (2007) *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*, Technical Report 07-49, University of Massachusetts, Amherst.

Huang, H., Li, Z., He, R., Sun, Z. and Tan, T. (2018a) *Introvae: Introspective Variational Autoencoders for Photographic Image Synthesis*, arXiv preprint arXiv:1807.06358.

Huang, H., Song, L., He, R., Sun, Z. and Tan, T. (2018b) *Variational Capsules for Image Analysis and Synthesis*, arXiv preprint arXiv:1807.04099.

Huang, W., Tu, S. and Xu, L. (2023) 'IA-faces: a bidirectional method for semantic face editing', *Neural Networks*, Vol. 158, pp.272–292.

Isola, P., Zhu, J-Y., Zhou, T. and Efros, A.A. (2017) 'Image-to-image translation with conditional adversarial networks', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1125–1134.

Kärkkäinen, K. and Joo, J. (2019) *Fairface: Face Attribute Dataset for Balanced Race, Gender, and Age*, arXiv preprint arXiv:1908.04913.

Karras, T., Aila, T., Laine, S. and Lehtinen, J. (2017) *Progressive Growing of GANs for Improved Quality, Stability, and Variation*, arXiv preprint arXiv:1710.10196.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J. and Aila, T. (2020) 'Training generative adversarial networks with limited data', *Advances in Neural Information Processing Systems*, Vol. 33, pp.12104–12114.

Karras, T., Laine, S. and Aila, T. (2019) 'A style-based generator architecture for generative adversarial networks', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.4401–4410.

Kazemi, H., Iranmanesh, M., Dabouei, A., Soleymani, S. and Nasrabadi, N.M. (2018) 'Facial attributes guided deep sketch-to-photo synthesis', in *2018 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, IEEE, pp.1–8.

Kingma, D.P. and Welling, M. (2013) *Auto-Encoding Variational Bayes*, arXiv preprint arXiv:1312.6114.

Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A. and Jain, A.K. (2015) 'Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus benchmark A', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1931–1939.

Ko, J., Cho, K., Choi, D., Ryoo, K. and Kim, S. (2023) '3D GAN inversion with pose optimization', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.2967–2976.

Kowalski, M., Garbin, S.J., Estellers, V., Baltrušaitis, T., Johnson, M. and Shotton, J. (2020) 'Config: controllable neural face image generation', in *European Conference on Computer Vision*, Springer, pp.299–315.

Kumar, N., Belhumeur, P. and Nayar, S. (2008) 'FaceTracer: a search engine for large collections of images with faces', in *European Conference on Computer Vision*, Springer, pp.340–353.

Kumar, N., Berg, A.C., Belhumeur, P.N. and Nayar, S.K. (2009) 'Attribute and simile classifiers for face verification', in *2009 IEEE 12th International Conference on Computer Vision*, IEEE, pp.365–372.

Kwak, J-g., Han, D.K. and Ko, H. (2020) 'Cafe-GAN: arbitrary face attribute editing with complementary attention feature', in *Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XIV 16*, Springer, Glasgow, UK, 23–28 August, pp.524–540.

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L. and Ranzato, M. (2017) *Fader Networks: Manipulating Images by Sliding Attributes*, arXiv preprint arXiv:1706.00409.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D.H., Hawk, S.T. and Van Knippenberg, A. (2010) 'Presentation and validation of the radboud faces database', *Cognition and Emotion*, Vol. 24, No. 8, pp.1377–1388.

Lee, C-H., Liu, Z., Wu, L. and Luo, P. (2020) 'MaskGAN: towards diverse and interactive facial image manipulation', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.5549–5558.

Li, D., Zhang, M., Zhang, L., Chen, W. and Feng, G. (2021) 'A novel attribute-based generation architecture for facial image editing', *Multimedia Tools and Applications*, Vol. 80, No. 4, pp.4881–4902.

Li, Q. and Tu, T. (2023) 'Large-pose facial makeup transfer based on generative adversarial network combined face alignment and face parsing', *Mathematical Biosciences and Engineering*, Vol. 20, No. 1, pp.737–757.

Li, L., Bao, J., Yang, H., Chen, D. and Wen, F. (2019) *Faceshifter: Towards High Fidelity and Occlusion Aware Face Swapping*, arXiv preprint arXiv:1912.13457.

Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W. and Lin, L. (2018) 'BeautyGAN: instance-level facial makeup transfer with deep generative adversarial network', in *Proceedings of the 26th ACM International Conference on Multimedia*, pp.645–653.

Li, B., Cai, S., Liu, W., Zhang, P., He, Q., Hua, M. and Yi, Z. (2023a) 'Dystyle: dynamic neural network for multi-attribute-conditioned style editings', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.189–197.

Li, S., Liu, L., Liu, J., Song, W., Hao, A. and Qin, H. (2023b) 'SC-GAN: subspace clustering based GAN for automatic expression manipulation', *Pattern Recognition*, Vol. 134, p.109072.

Ling, J., Xue, H., Song, L., Yang, S., Xie, R. and Gu, X. (2020) 'Toward fine-grained facial expression manipulation', in *European Conference on Computer Vision*, Springer, pp.37–53.

Liu, H., Xu, Y. and Chen, F. (2023) 'Sketch2Photo: synthesizing photo-realistic images from sketches via global contexts', *Engineering Applications of Artificial Intelligence*, Vol. 117, p.105608.

Liu, M., Ding, Y., Xia, M., Liu, X., Ding, E., Zuo, W. and Wen, S. (2019) 'STGAN: a unified selective transfer network for arbitrary image attribute editing', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.3673–3682.

Liu, S., Sun, Y., Zhu, D., Bao, R., Wang, W., Shu, X. and Yan, S. (2017) 'Face aging with contextual generative adversarial nets', in *Proceedings of the 25th ACM International Conference on Multimedia*, pp.82–90.

Liu, Z., Luo, P., Wang, X. and Tang, X. (2015) 'Deep learning face attributes in the wild', in *Proceedings of the IEEE International Conference on Computer Vision*, ppp.3730–3738.

Luo, X., Chen, X., He, X., Qing, L. and Tan, X. (2022) 'CMAFGAN: a cross-modal attention fusion based generative adversarial network for attribute word-to-face synthesis', *Knowledge-Based Systems*, Vol. 255, p.109750.

Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T. and Van Gool, L. (2018) *Exemplar Guided Unsupervised Image-to-Image Translation with Semantic Consistency*, arXiv preprint arXiv:1805.11145.

Mirza, M. and Osindero, S. (2014) *Conditional Generative Adversarial Nets*, arXiv preprint arXiv:1411.1784.

Mollahosseini, A., Hasani, B. and Mahoor, M.H. (2017) 'AffectNet: a database for facial expression, valence, and arousal computing in the wild', *IEEE Transactions on Affective Computing*, Vol. 10, No. 1, pp.18–31.

Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I. and Zafeiriou, S. (2017) 'AgeDB: the first manually collected, in-the-wild age database', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp.51–59.

Ning, X., Xu, S., Li, W. and Nie, S. (2020) 'FEGAN: flexible and efficient face editing with pre-trained generator', *IEEE Access*, Vol. 8, pp.65340–65350.

Ning, X., Li, W., Dong, X., Xu, S., Nan, F. and Yao, Y. (2021) 'Continuous learning of face attribute synthesis', in *2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, pp.4282–4289.

Parihar, R., Dhiman, A., Karmali, T. and Babu, R.V. (2022) *Everything is there in Latent Space: Attribute Editing and Attribute Style Manipulation by StyleGAN Latent Space Exploration*, arXiv preprint arXiv:2207.09855.

Parkhi, O.M., Vedaldi, A. and Zisserman, A. (2015) *Deep Face Recognition*, British Machine Vision Association.

Phillips, P.J., Moon, H., Rizvi, S.A. and Rauss, P.J. (2000) 'The FERET evaluation methodology for face-recognition algorithms', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 10, pp.1090–1104.

Phusomsai, W. and Limpiyakorn, Y. (2020) 'Applying GANs for generating image with varied facial attributes from sketch', in *Journal of Physics: Conference Series*, IOP Publishing, Vol. 1619, p.p12013.

Ranjan, R., Patel, V.M. and Chellappa, R. (2017) 'Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 1, pp.121–135.

Roich, D., Mokady, R., Bermano, A.H. and Cohen-Or, D. (2022) 'Pivotal tuning for latent-based editing of real images', *ACM Transactions on Graphics (TOG)*, Vol. 42, No. 1, pp.1–13.

Romero, A., Van Gool, L. and Timofte, R. (2021) 'Smile: semantically-guided multi-attribute image and layout editing', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.1924–1933.

Rothe, R., Timofte, R. and Van Gool, L. (2018) 'Deep expectation of real and apparent age from a single image without facial landmarks', *International Journal of Computer Vision*, Vol. 126, No. 2, pp.144–157.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X. (2016) 'Improved techniques for training GANs', *Advances in Neural Information Processing Systems*, Vol. 29.

Sengupta, S., Chen, J-C., Castillo, C., Patel, V.M., Chellappa, R. and Jacobs, D.W. (2016) 'Frontal to profile face verification in the wild', in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp.1–9.

Shen, W. and Liu, R. (2017) 'Learning residual images for face attribute manipulation', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.4030–4038.

Shi, R-x., Ye, D-y. and Chen, Z-j. (2022) 'A bi-directional facial attribute transfer framework: transfer your single facial attribute to a portrait illustration', *Neural Computing and Applications*, Vol. 34, No. 1, pp.253–270.

Shiri, F., Yu, X., Porikli, F., Hartley, R. and Koniusz, P. (2019) 'Recovering faces from portraits with auxiliary facial attributes', in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp.406–415.

Song, X., Shao, M., Zuo, W. and Li, C. (2020) 'Face attribute editing based on generative adversarial networks', *Signal, Image and Video Processing*, Vol. 14, No. 6, pp.1217–1225.

Sun, R., Huang, C., Zhu, H. and Ma, L. (2021) 'Mask-aware photorealistic facial attribute manipulation', *Computational Visual Media*, pp.1–12.

Sun, Q., Guo, J. and Liu, Y. (2022) 'PattGAN: pluralistic facial attribute editing', *IEEE Access*, Vol. 10, pp.68534–68544.

Szegedy, C., Ioffe, S., Vanhoucke, V. and Alemi, A.A. (2017) 'Inception-V4, inception-resnet and the impact of residual connections on learning', in *Thirty-First AAAI Conference on Artificial Intelligence*.

Tan, Z., Chai, M., Chen, D., Liao, J., Chu, Q., Yuan, L., Tulyakov, S. and Yu, N. (2020) *MichiGAN: Multi-Input-Conditioned Hair Image Generation for Portrait Editing*, arXiv preprint arXiv:2010.16417.

Tewari, A., Elgharib, M., Bernard, F., Seidel, H-P., Pérez, P., Zollhöfer, M. and Theobalt, C. (2020) 'PIE: portrait image embedding for semantic control', *ACM Transactions on Graphics (TOG)*, Vol. 39, No. 6, pp.1–14.

Viazovetskyi, Y., Ivashkin, V. and Kashin, E. (2020) 'StyleGAN2 distillation for feed-forward image manipulation', in *European Conference on Computer Vision*, Springer, pp.170–186.

Wang, F., Chen, L., Li, C., Huang, S., Chen, Y., Qian, C. and Loy, C.C. (2018a) 'The devil of face recognition is in the noise', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.765–780.

Wang, Y., Wang, S., Qi, G., Tang, J. and Li, B. (2018b) 'Weakly supervised facial attribute manipulation via deep adversarial network', in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, pp.112–121.

Wang, F., Xiang, S., Liu, T. and Fu, Y. (2021) 'Attention based facial expression manipulation', in *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, pp.1–6.

Wang, Z., Bovik, A.C., Sheikh, H.R. and Simoncelli, E.P. (2004) 'Image quality assessment: from error visibility to structural similarity', *IEEE Transactions on Image Processing*, Vol. 13, No. 4, pp.600–612.

Wei, Y., Gan, Z., Li, W., Lyu, S., Chang, M-C., Zhang, L., Gao, J. and Zhang, P. (2020) 'MagGAN: high-resolution face attribute editing with mask-guided generative adversarial network', in *Proceedings of the Asian Conference on Computer Vision*.

Wolf, L., Hassner, T. and Taigman, Y. (2010) 'Effective unconstrained face recognition by combining multiple descriptors and learned background statistics', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 33, No. 10, pp.1978–1990.

Wolf, L., Hassner, T. and Maoz, I. (2011) 'Face recognition in unconstrained videos with matched background similarity', in *CVPR 2011*, IEEE, pp.529–534.

Wu, H., He, F., Si, T., Duan, Y. and Yan, X. (2023) 'HIGSA: human image generation with self-attention', *Advanced Engineering Informatics*, Vol. 55, p.101856.

Xiao, T., Hong, J. and Ma, J. (2018) 'Elegant: exchanging latent encodings with GAN for transferring multiple face attributes', in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.168–184.

Yang, B., Chen, X., Hong, R., Chen, Z., Li, Y. and Zha, Z-J. (2020a) 'Joint sketch-attribute learning for fine-grained face synthesis', in *International Conference on Multimedia Modeling*, Springer, pp.790–801.

Yang, Y., Hossain, M.Z., Gedeon, T. and Rahman, S. (2020b) *S2FGAN: Semantically Aware Interactive Sketch-to-Face Translation*, arXiv preprint arXiv:2011.14785.

Yang, S., Wang, Z., Liu, J. and Guo, Z. (2021) 'Controllable sketch-to-image translation for robust face synthesis', *IEEE Transactions on Image Processing*, Vol. 30, pp.8797–8810.

Yang, S., Jiang, L., Liu, Z. and Loy, C.C. (2022) 'Pastiche master: exemplar-based high-resolution portrait style transfer', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.7693–7702.

Yin, W., Liu, Z. and Loy, C.C. (2019) 'Instance-level facial attributes transfer with geometry-aware flow', in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp.9111–9118.

Ying, L., Fan, H., Ni, F. and Xiang, J. (2019) *CLSGAN: Selective Attribute Editing based on Classification Adversarial Network*, arXiv preprint arXiv:1910.11764.

Yoo, S-M., Choi, T-M., Choi, J-W. and Kim, J-H. (2023) 'FastSwap: a lightweight one-stage framework for real-time face swapping', in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp.3558–3567.

Zhang, Z., Song, Y. and Qi, H. (2017) 'Age progression/regression by conditional adversarial autoencoder', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.5810–5818.

Zhang, J., Li, A., Liu, Y. and Wang, M. (2019) 'Adversarially regularized U-Net-based GANs for facial attribute modification and generation', *IEEE Access*, Vol. 7, pp.86453–86462.

Zhang, K., Su, Y., Guo, X., Qi, L. and Zhao, Z. (2020) 'MU-GAN: facial attribute editing based on multi-attention mechanism', *IEEE/CAA Journal of Automatica Sinica*, Vol. 8, No. 9, pp.1614–1626.

Zhou, S., Xiao, T., Yang, Y., Feng, D., He, Q. and He, W. (2017a) *GeneGAN: Learning Object Transfiguration and Attribute Subspace from Unpaired Data*, arXiv preprint arXiv:1705.04932.

Zhou, T., Brown, M., Snavely, N. and Lowe, D.G. (2017b) ‘Unsupervised learning of depth and ego-motion from video’, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.1851–1858.

Zhu, D., Liu, S., Jiang, W., Gao, C., Wu, T., Wang, Q., and Guo, G. (2019a) *UGAN: Untraceable GAN for Multi-Domain Face Translation*, arXiv preprint arXiv:1907.11418.

Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B. and Bai, X. (2019b) ‘Progressive pose attention transfer for person image generation’, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.2347–2356.