Noname manuscript No.
(will be inserted by the editor)

A Survey to Deep Facial Attribute Analysis

Xin Zheng · Yanqing Guo* · Huaibo Huang · Yi Li · Ran He ·

Received: date / Accepted: date

Abstract Facial attribute analysis has received considerable attention when deep learning techniques make remarkable breakthroughs in this field over the past few years. Deep learning based facial attribute analysis is comprised of two basic sub-issues: Facial Attribute Estimation (FAE), which recognizes whether facial attributes are present in given images, and Facial Attribute Manipulation (FAM), which synthesizes or removes desired facial attributes. In this paper, we provide a comprehensive survey on deep facial attribute analysis from the perspectives of both estimation and manipulation. First, we summarize that deep facial attribute analysis follows a general pipeline, which comprises two stages, i.e., data pre-processing and model construction. Meanwhile, the underlying theories of the two-stage pipeline are provided for both FAE and FAM, respectively. Second, we introduce commonly used datasets and performance metrics in facial attribute analysis. Third, we create a taxonomy of the state-of-the-arts and review FAE and FAM algorithms in detail. Furthermore, several additional facial attribute related issues are introduced, as well as relevant real-world applications. Finally, we discuss possible challenges and promising future research directions.

Keywords Deep Neural Networks · Deep Facial Attribute Analysis · Facial Attribute Estimation · Facial Attribute Manipulation

Xin Zheng
School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
E-mail: zhengxin@mail.dlut.edu.cn

Yanqing Guo*
School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China
E-mail: guoyq@dlut.edu.cn

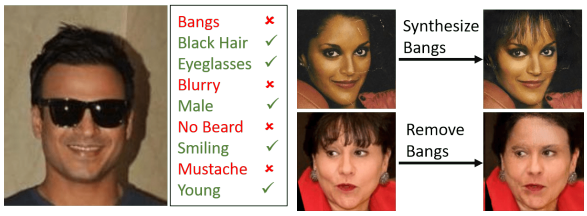
Huaibo Huang
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100190, China
E-mail: huaibo.huang@cripac.ia.ac.cn

Yi Li
National Laboratory of Pattern Recognition, CASIA
Center for Research on Intelligent Perception and Computing, CASIA
University of Chinese Academy of Sciences, Beijing 100190, China E-mail: yi.li@cripac.ia.ac.cn

Ran He
National Laboratory of Pattern Recognition, CASIA
Center for Research on Intelligent Perception and Computing, CASIA
Center for Excellence in Brain Science and Intelligence Technology, CAS
University of Chinese Academy of Sciences, Beijing 100190, China E-mail: rhe@nlpr.ia.ac.cn

1 Introduction

As the vital information of faces, facial attributes contribute to numerous successful real-world applications, e.g., face verification [54,3,97,119,10], face recognition [40,94,39,100,83], face retrieval [63,76,22,106] and face image synthesis [47,7,48,98,20]. In general, facial attributes represent intuitive semantic features that describe human-understandable visual properties of face images, such as *smiling*, *eyeglasses*, *mustache*, etc. Facial attribute analysis, aiming to build a bridge between such human-understandable visual descriptions and abstract feature presentations required by many computer vision tasks, has attracted growing attention and become a hot research topic. Recently, the development of deep learning techniques makes excellent progress in the learning of abstract feature presentations, leading to significant performance improvement of current algorithms in the field of deep facial attribute analysis.



(a) FAE (b) FAM

Fig. 1: The illustration of two sub-issues in deep facial attribute analysis (i.e., FAE and FAM).

Deep facial attribute analysis mainly contains two sub-issues: Facial Attribute Estimation (FAE) and Facial Attribute Manipulation (FAM). Given a face image, FAE recognizes whether a describable attribute of visual appearance is present by training attribute classifiers. In contrast, FAM modifies face images to synthesize or remove desired attributes by constructing generative models. The illustration of both FAE and FAM is provided in Fig. 1.

Deep facial attribute estimation (FAE) methods can be generally categorized into two groups: part-based methods and holistic methods. Part-based FAE methods first locate the positions of facial attributes and further extract features according to the obtained localization cues for the subsequent attribute prediction. Depending on different schemes of locating facial attributes, part-based methods can be further classified into two sub-categories: separate auxiliary localization methods and end-to-end localization counterparts. More specifically, separate auxiliary localization ones seek help from existing part detectors or auxiliary localization algorithms, e.g., facial key point detection [72,111] and semantic segmentation [49,25], and then learn features from different positions for further estimation. Note that the localization and estimation are operated in a separate and independent manner. In contrast, end-to-end localization methods exploit the locations of facial attributes and predict their presences simultaneously in end-to-end frameworks. Compared with part-based methods, holistic methods focus more on learning attribute relationships and estimating facial attributes in a unified framework without any extra localization modules. Specifically, such methods model the association and distinction among different attributes to explore the complementary information, by designing various networks with sharing features from different layers. Besides, other prior or auxiliary information, such as attribute-group partition or identity information assistance [8], are also taken into consideration when predicting attributes in the holistic frameworks.

Deep facial attribute manipulation (FAM) methods are constructed mainly based on generative models, of which generative adversarial networks (GANs) and variational autoencoders (VAEs) play as the backbones. Further, these FAM algorithms can be grouped into two categories: model-based methods and extra condition-based methods, where the critical difference between them is whether extra conditions are required. Model-based methods construct a model without any extra conditional inputs and learn a set of model parameters that only correspond to one attribute during a single training process. That means when editing another attribute, another training process needs to be executed in the same way. In this case, multiple attribute manipulations correspond to multiple training processes, resulting in expensive calculation costs. In contrast, extra condition-based methods take extra attribute vectors or reference images as input conditions, which can alter multiple attributes simultaneously by changing the corresponding values of attribute vectors or taking multiple exemplars with distinct attributes as references. Specifically, given the original to-be-manipulated images, extra conditional attribute vectors, such as one-hot vectors indicating the attribute presence, are concatenated with the latent original image codes. By comparison, extra conditional reference exemplars exchange specific attributes with the original images in the framework of the image-to-image translation. Note that these reference images do not need to have the same identity with the original to-be-manipulate images. Due to more abundant facial details and more photo-realistic results caused by attribute transfer based on reference images, recently, more and more researchers shift their focus on such methods to generate more faithful facial attribute images [124,113,71]. In this way, much more specific details of references can be discovered rather than merely altering the values of attribute vectors to edit facial attributes.

To sum up, we clarify the whole diverse categories of deep facial attribute analysis in a tree diagram as shown in Fig. 2. Further, aiming at summarizing the progress in deep facial attribute analysis, milestones of both facial attribute estimation and manipulation are listed in Fig. 3 and in Fig. 4, respectively.

As shown in Fig. 3, two types of deep facial attribute estimation methods, i.e., part-based ones and holistic counterparts, share two parallel routes. The study of deep FAE can be traced back to the earliest part-based work of Zhang et al. [118], who take the whole person image as input and partition face related regions aiming at estimating several facial attributes. In the meantime, torso related regions are also explored to predict other human body attributes. Then, the emer-

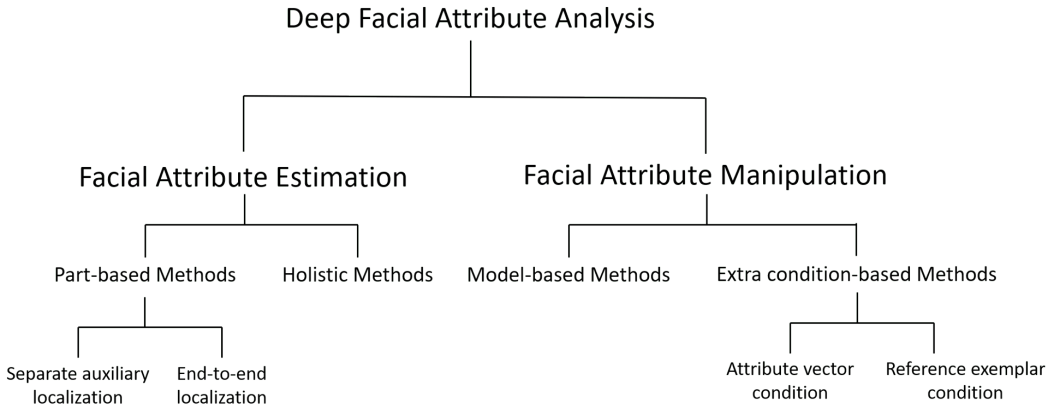


Fig. 2: Tree diagram for diverse categories of deep facial attribute analysis algorithms.

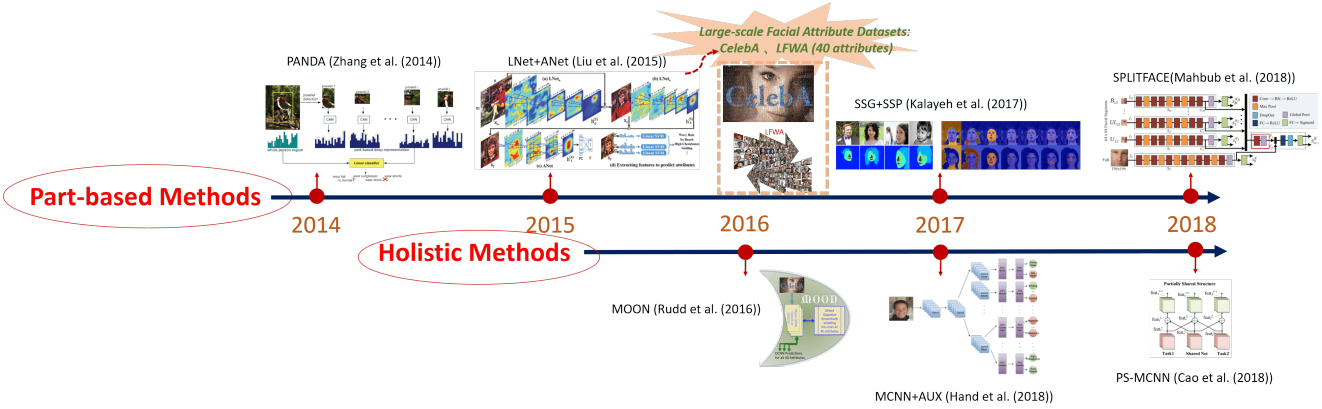


Fig. 3: The evolution of facial attribute estimation (FAE) methods.
(Best viewed by zooming in the electronic version)

gence of LNet+ANet [66] pushes deep FAE into an independent research branch, where only face images are taken as inputs for merely estimating face related attributes. Besides, two large-scale datasets, i.e., CelebA and LFWA, with 40 labeled facial attributes, are released and further contribute to the development of deep FAE methods. After this, part-based methods and holistic methods share common development and success but have distinct directions and trends. Part-based methods place more emphasis on facial details for modeling localization cues, whereas holistic methods are inclined to design more specific networks and learn more discriminative features with attribute relationships.

The development of deep FAM methods is illustrated in Fig. 4. Model-based methods and two kinds of extra condition-based methods follow their evolution processes respectively, but all pursue the development of GANs [26,75,12] and VAEs [51,17,47,48]. The earliest deep FAM work DIAT [61], a model-based method, first attempts to utilize simple GANs to generate facial attributes. In the meantime, the development of the conditional GAN and VAE brings an opportunity for extra condition-based methods, especially those conditioned on attribute vectors, to dominate deep FAM due to their advantages of changing multiple attributes simultaneously. However, extra attribute vector-based methods cannot guarantee the rest details that are irrelevant to manipulated attributes keep unchanged, while model-based methods can conquer this problem. Until now, how to address the balance between changing multiple interested attributes and preserving other irrelevant ones concurrently is an open issue to be solved. In light of this, methods conditioned on reference exemplars are coming into researchers' insights for tackling such challenge to generate more and more photo-realistic facial attribute images. We have reason to believe that exemplar-guided facial attribute manipulation methods will become a popular research trend.

Although a large number of methods achieve appealing performance in deep FAE and FAM, there still

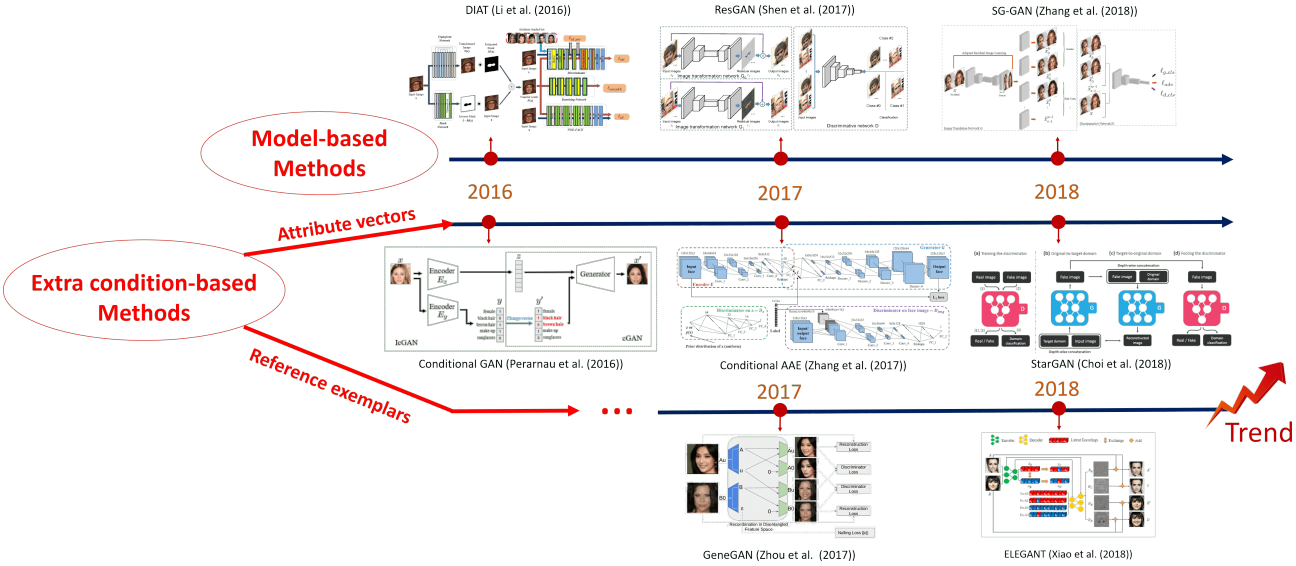


Fig. 4: The evolution of facial attribute manipulation (FAM) methods. (Best viewed by zooming in the electronic version)

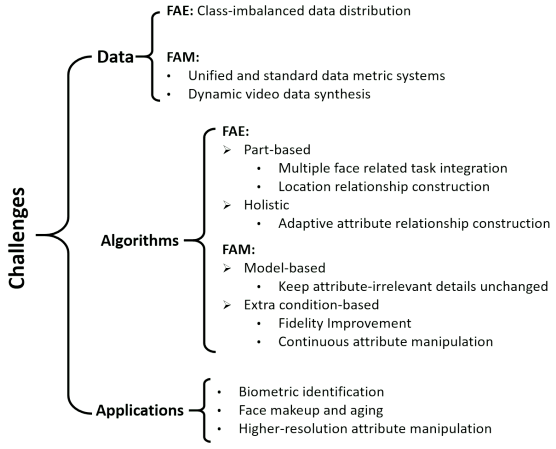


Fig. 5: Summary of challenges in deep facial attribute analysis.

remain several challenges in facial attribute analysis. Generally, these core challenges can be summarized into the following aspects: data, algorithms, and applications, as illustrated in Fig. 5.

First, from the perspective of data, deep FAE methods eagerly require more diverse data sources, when existing datasets have the problem with significantly class-imbalanced distribution. That means some of the classes have a much higher number of examples than others, corresponding to the majority class and minority class, respectively. As a result, the learned model would bias towards the majority samples and perform poorly in the real-world data. As for the data challenge faced by deep FAM methods, seeking unified and standard data metric systems concerning qualitative and quantitative analysis is a thorny problem. Since a large proportion of current algorithms evaluate the quality of their generated images based on the visual fidelity, such measurement lacks agreed protocols and harms the performance evaluation. Besides, recently, the rapid development of multi-media in the era of big data has given birth to rich video data. This case leads to more tough challenges for deep FAM algorithms when the significant difficulty exists in the synthesis of dynamic data.

Second, looking at the details, as regards algorithms, deep part-based FAE methods mainly focus on two aspects. The first is attempting to integrate multiple face related tasks, such as landmark detection and face detection, into a unified framework along with attribute estimation. In this way, the complementary information among different tasks can be fully exploited. For the second aspect, the in-the-wild data contain various illumination variations, occlusions, and non-frontal faces caused by environmental conditions, which significantly degrade the performance of existing deep FAE algorithms trained over laboratory data. Hence, part-based FAE methods are expected to pay more attention to discover the relationship among different attribute locations in the future. As for the deep holistic FAE methods, the critical challenge in algorithms exists in adaptive attribute relationship construction, when almost current methods model attribute relationships with the help of the prior information, such as human-made attribute groups. Such artificially subjec-

tive attribute partition may not be optimal, and more adaptive schemes need to be lighted on. When it comes to the challenges faced by deep FAM methods, model-based methods have a serious drawback: they cannot keep the other attribute-irrelevant details unchanged as the supervised information merely comes from the target images with desired attributes. Besides, extra condition-based methods need to work harder on manipulating attributes continuously, where the strategy of interpolation is expected to be the solution to this case. Moreover, as we mentioned above, reference exemplar-based FAM will be a popular research direction for generating more faithful and photo-realistic images in the future.

Finally, when it comes to the challenging applications, biometric identification has become an essential technique to protect the private information of users in lots of mobile devices, where biological characteristics, such as fingerprints and irises, are used to avoid various attacks targeting for personal details. Facial attributes, which contain much more detailed characteristics about faces, can be significantly considered as a rising biological characteristic and begin to attract the attention of researchers in the field of mobile device authentication. Moreover, face makeup and face aging are hot topics for face recognition and entertainment related applications in deep FAM. In face makeup, the *heavy makeup* attribute branches out more subtle subcategories by changing lip colors, the shapes of eyebrows, and accentuating eye regions. As for the face aging, more distinct age groups are partitioned in a wide range of ages, so that a manipulated face can be synthesized whose target age lies in some given group. Furthermore, note that the severe restriction of current deep FAM algorithms is that they only work well at a limited range of resolution. Such restriction becomes another challenge to operate on face images with higher resolutions for deep FAM methods in the real-world applications.

Despite the tough challenges faced by both FAE and FAM, the complementary relationships between the two might bring us an opportunity to make both better. On the one hand, facial attribute manipulation can be taken as a vital scheme of data augmentation for facial attribute estimation, where generated facial attribute images can significantly increase the amount of data so as to relieve the overfitting further. On the other hand, facial attribute estimation can be a significant quantitative performance evaluation criterion of facial attribute manipulation, where the accuracy gap between real images and generated images can be used to reflect the performance of FAM algorithms.

In this paper, we make an in-depth survey of facial attribute analysis based on deep learning, including FAE and FAM. The primary goal is to provide an overview of the two issues, as well as point out their respective strengths and weaknesses to give a newcomer prime skills for this field. The rest of this paper is organized as follows. In Section 2, we summarize deep facial attribute analysis into a general two-stage pipeline, including data pre-processing and model construction. Then, we further provide the underlying theories of both FAE and FAM in such pipeline, respectively. In Section 3, we summarize commonly used publicly facial attribute datasets and metrics. Section 4 and Section 5 provide detailed overviews of state-of-the-art deep learning based FAE and FAM methods, as well as their pros and cons, respectively. Additional related issues and challenges are discussed in Section 6 and Section 7, respectively. Finally, we conclude this paper in Section 8.

2 Facial Attribute Analysis Preliminaries

We summarize deep facial attribute analysis as a general pipeline that comprises two stages: data pre-processing and model construction, as shown in Fig. 6. In this section, first, we introduce two commonly used data pre-processing strategies for both FAE and FAM, including face detection and alignment, as well as data augmentation. Second, we provide the general processes of the model construction for both FAE and FAM in detail, respectively. Specifically, for deep FAE methods, we provide the basics about feature extraction and attribute classification, which are two crucial parts when designing facial attribute estimation models. For deep FAM methods, we review the underlying theories of VAEs and GANs, as well as their conditional versions.

2.1 Data Pre-processing

Given a large-scale facial attribute dataset, the variations of illuminations, poses, occlusions, and low image qualities are primary influence factors in unconstrained scenarios for deep facial attribute analysis. Therefore, data pre-processing is usually the first and necessary step before training deep networks for both FAE and FAM, where face detection and alignment, as well as data augmentation, are two commonly used schemes.

2.1.1 Face Detection and Alignment

Before the databases with more facial attribute annotations are released, most of attribute prediction methods [118,52,24] must take the whole human images (more than face regions) as inputs, and only several well-marked

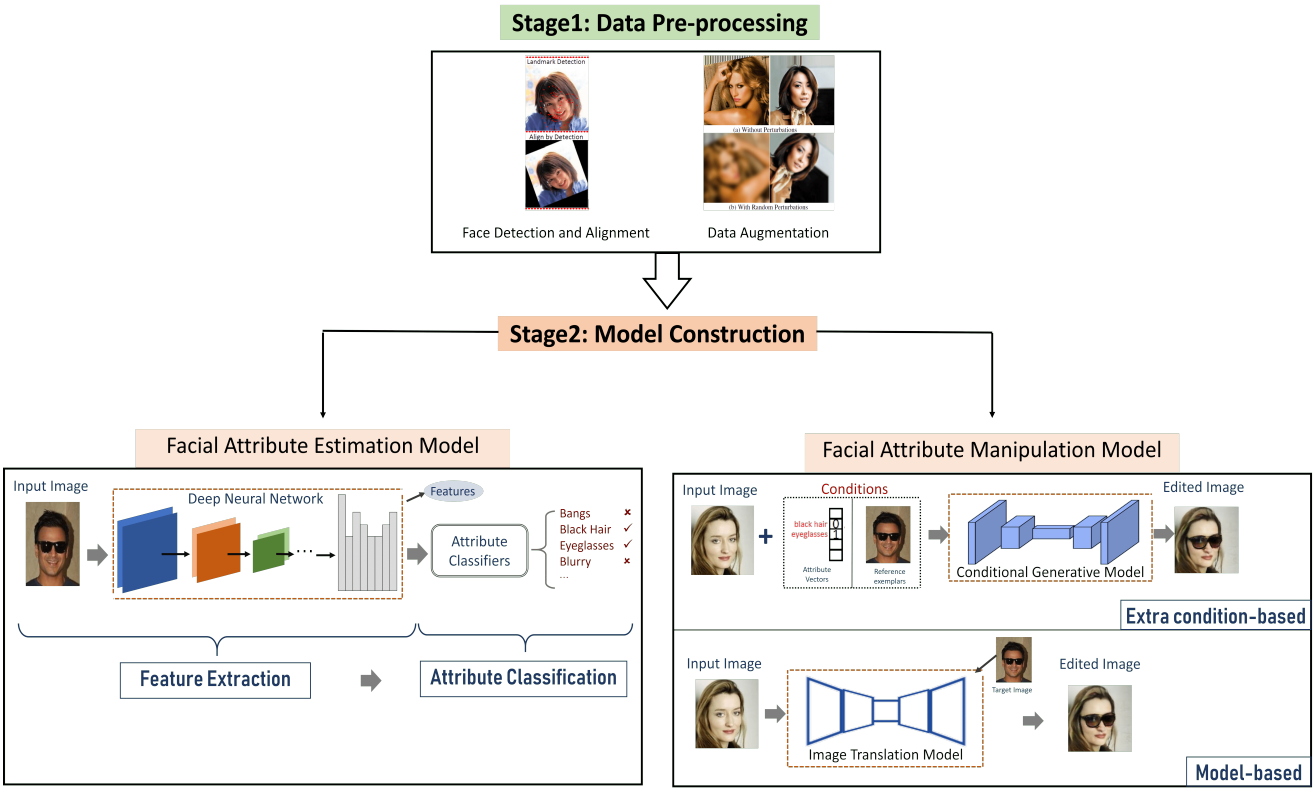


Fig. 6: Two-stage pipeline of deep facial attribute analysis.

facial attributes can be estimated, such as *smile*, *gender*, and *has glasses*. There is no doubt that much extra irrelevant information is involved, resulting in redundant computations. Hence, face detection and alignment become crucial steps to locate face areas for reducing the calculation of attribute-irrelevant information.

In face detection, Ranjan et al. [81] first recognize the *gender* attribute with a HyperFace detector that locates faces and landmarks, and then, Günther et al. [29] further extend to predict 40 facial attributes simultaneously with the same HyperFace. In contrast, Kumar et al. [52] use a poselet part detector [6] to detect different parts corresponding to different poses, where the face is taken as a part of an image of the whole person region. Compared with such poselet detector operated over gradient orientation features, Gkioxari et al. [24] propose a 'deep' version of the poselet, which trains a sliding window detector operated on deep feature pyramids. Specifically, the deep poselet detector divides the human body into three parts (head, torso, and legs) and clusters fiducial key points of each part into many different poselets. However, as all the mentioned face detectors are used to find the rough facial part and other body parts, more subtle facial attributes related to details, such as *eyebrows*, cannot be well predicted.

In face alignment, well-annotated databases with fiducial key points have significant benefits for both FAE and FAM methods, since more specific facial regions of attributes can be captured expediently. Consequently, unalignment error would not make features corresponding to specific regions go outside of these parts. All-in-One Face dataset [82] provides fiducial key points and full faces. Based on these fiducial key points, Mahbub et al. [72] divide a face into 14 facial segments related to different facial regions, working on the attribute prediction of partial faces. Kumar et al. [52] artificially break up the face into 10 functional parts including hair, forehead, eyebrows, eyes, nose, cheeks, upper lip, mouth, and chin. Such partition ensures that these regions of the face are wide enough to be robust against discrepancies among individual faces and overlap slightly. In this way, the geometry shared by different faces can be exploited, and small errors in alignment would not degenerate the performance of features.

Recently, there is a trend integrating face detection and alignment into the training process in facial attribute analysis. He et al. [36] take the face detection as a special case of the general semi-rigid object detection. Consequently, the performance of both face detection and attribute estimation is improved through the jointly learned architecture. More importantly, the

input images from the wild, which vary a lot in illuminations and occlusions, can also be well handled in this way without cropping and aligning. Ding et al. [16] propose a cascade network to localize the face regions according to different attributes and perform facial attribute estimation simultaneously with no need to align faces. Li et al. [60] design an AFFAIR network for learning a hierarchy of spatial transformation and predicting facial attributes without landmarks.

To sum up, more and more researches are devoted to integrating the face detection and alignment into the training process, even exploring schemes to refrain from face detection and alignment.

2.1.2 Data Augmentation

In most face processing tasks, data augmentation is a vital strategy for solving the problems of insufficient training data and overfitting in deep learning, and face attribute analysis is no exception. By imposing perturbations and distortions on the input images, data can be extended for improving deep learning models.

Günther et al. [29] propose an Alignment-Free Facial Attribute Classification Technique (AFFACT) with data augmentation. More specifically, AFFACT leverages shifts, rotations, and scales of images to make facial attribute feature extraction more reliable in both training stage and testing stage. In the training stage, first, face images are rotated, scaled, cropped, and horizontally flipped with 50% probability with defined coordinates. Then, a Gaussian filter is applied to emulate smaller image resolutions and yield blurred upscaled images. In the testing stage, AFFACT rescales the test images at first. Then, it transforms the images to 10 crops including a center one, four corners of the original images, as well as their horizontally flipped versions. Finally, it averages the scores from the deep network per attribute over these ten crops to make the final prediction. Apart from taking crops, AFFACT also uses all combinations of shifts, scales, and angles, as well as their mirrored versions. Consequently, AFFACT enhances the performance of the deep network in FAE effectively.

2.2 Basis of FAE Model Construction

The main difference between two types of deep FAE methods, i.e., part-based methods and holistic methods, lies in the stage of deep neural network model construction. Specifically, FAE model construction generally contains two phases: feature extraction and attribute classification, as shown in Fig 6. Deep neural networks are designed to extract deep features, so as to estimate the presences of facial attributes by attribute classifiers in the following step. The same goal can be achieved in an end-to-end way as well, where feature learning and attribute classification are implemented in a unified framework. We provided more details about the basis of FAM model construction as follows.

2.2.1 Feature Extraction

Deep convolutional neural networks (CNNs) play significant roles in learning discriminative representations and achieve state-of-the-art performance for deep facial attribute estimation. In general, arbitrary classical CNN networks which learn to focus on detailed properties, such as VGG [78] and ResNet [38], can be used to extract deep facial attribute features. Zhong et al. [122] directly apply the FaceNet and VGG-16 network for learning facial features, and Günther et al. [29] investigate the performance of the ResNet.

To analyze how the features at different levels of networks affect the FAE performance, Zhong et al. [123] take mid-level CNN features as an alternative to the high-level ones for attribute prediction. The experiments demonstrate that even the early convolution layers astonishingly achieve comparable performance in most facial attributes as the state-of-the-arts, and such mid-level representations are capable of breaking the bounds of the interconnections between convolutional and fully connected (FC) layers, resulting in accepted arbitrary receptive sizes of CNNs.

According to the two types of deep FAE methods, i.e., part-based methods and holistic methods, researchers consider the following two problems when designing networks for discriminative feature extraction.

(1) how to make networks focus more on the locations of attributes?

(2) how to make the best of the relationships among attributes?

We provide more details for the two concerns when introducing state-of-the-art methods hereinafter.

Besides, there exist methods designing specific network architectures for learning discriminative features. Lu et al. [67] design an automatically constructed compact multi-task deep learning architecture, which starts with a thin multi-layer network and dynamically widens in a greedy manner. Belghazi et al. [2] build a hierarchical generative model and a corresponding inference model through the adversarial learning paradigm.

2.2.2 Attribute Classification

Early methods learn deep feature representations with deep networks but make predictions with traditional

classifiers, such as support vector machines (SVMs) [15, 5], decision trees [70], as well as k-nearest neighbor (kNN) algorithm [44, 45]. For example, Kumar et al. [54] train multiple SVMs [15] with radial basis function (RBF) kernels for the multiple attribute prediction, where each linear SVM corresponds to one facial attribute. Bourdev et al. [5] present a feedforward classification system with linear SVMs and classify attributes at the image patch level, the whole image level, and the semantic relationship level, respectively. Luo et al. [70] construct a sum-product decision tree network for the prediction of facial attributes. Besides, Huang et al. [44, 45] adopt the kNN algorithm for dealing with the class-imbalanced FAE problem.

Looking at details, as regards attribute classifiers based on deep networks, FC layers attached to the end of deep feature extraction networks achieve the purpose of the basic facial attribute estimation. However, what the most crucial is to measure the losses between the outputs of FC layers and the ground truths with proper loss functions for reducing the classification error.

Rudd et al. first take the multiple facial attribute classification as a regression issue to minimize the mean squared error (MSE) loss, i.e., the Euclidean loss, by mixing errors of all attributes. They train a mixed objective optimization network [86] (MOON) based on the VGG-16 topology [78]. In this way, multiple attribute labels can be obtained simultaneously via a single deep convolutional neural network (DCNN). Rozsa et al. [84] also adopt the Euclidean loss to train multiple DCNNs, where each one is for each facial attribute.

Apart from the Euclidean loss, the most commonly used loss function is the sigmoid cross entropy loss, which treats each attribute prediction as a binary classification task. Hand and Chellappa [34] develop a Multi-task deep CNN (MCNN) and use the sigmoid cross-entropy loss to evaluate its output, as well as calculate the scores of all facial attribute. Furthermore, the output scores are fed into an auxiliary network (AUX) for learning attribute correlations at the score level. Moreover, Günther et al. [29] even provide an evaluation of the Euclidean loss and the sigmoid cross entropy loss. The experiments over the same network but different loss functions denote that the two loss functions achieve comparable performance, which illustrates that the two loss functions do not affect much on the performance under the same network.

2.3 Basis of FAM Model Construction

As shown in Fig 6, when constructing facial attribute manipulation models, model-based methods and extra condition-based methods follow utterly different design principles. Model-based methods train a model only corresponding to one attribute, and the whole process follows an image translation paradigm, where the training is supervised by the target images with desired attributes to be synthesized or removed. Extra condition-based methods construct conditional generative models according to different conditions, i.e., attribute vectors or reference exemplars. However, almost all deep FAM models are constructed based on VAEs and GANs. Hence, we below review basic theories about VAEs and GANs, as well as their respective condition versions. More detailed algorithms about each type of current deep FAM methods are introduced in Section 5.

2.3.1 Variational autoencoder

In general, a VAE has two components: the encoder, which encodes a data x into a latent representation z, i.e., $z \sim Encoder(x) = q(z|x)$, and the decoder, which maps the obtained latent representation z back to the data space x, i.e., $\tilde{x} \sim Decoder(z) = p(x|z)$. VAE first samples z from the distribution of encoder $p(z)$, where $z \sim \mathcal{N}(0, I)$ typically. Then, the obtained sample is fed into the differentiable generator network $g(z)$ for obtaining $\tilde{x}$ by sampling from $p(x; g(z)) = p(x|z)$. The key of VAE is training to max the variational lower bound $\mathcal{L}_{VAE}(q)$ [47]:

$$\mathcal{L}_{VAE}(q) = \mathbb{E}_{z\sim q(z|x)} \log p\left(z, x\right) - D_{KL}\left(q\left(z|x\right) || p\left(z\right)\right), \tag{1}$$

where D_{KL} denotes the Kullback-Leibler divergence.

As for the conditional version of VAE, given the attribute vector y and latent representation z, it aims to build a model $p(x|y, z)$ for generating images x containing desired attributes, where taking y and z as conditional variables. Such image generation task follows a two-step process: the first step is randomly sampling latent variable z from prior distribution $p(z)$, while the second step is generating image according to given conditional variables. As a result, the variational lower bound of conditional VAE can be written as [114]

$$\mathcal{L}_{CVAE}(q) = \mathbb{E}_{z\sim q(z|x,y)} \log p\left(x|y, z\right) - D_{KL}\left(q\left(z|x, y\right) || p\left(z\right)\right), \tag{2}$$

where $q(z|x, y)$ is the true posterior from the encoder.

2.3.2 Generative adversarial network

A generative adversarial network (GAN) contains two parts: the generator G and the discriminator D, where G tries to synthesize data from a random vector z obeying a prior noise distribution $z \sim p\left(z\right)$, and D attempts

to discriminate that whether data are from realistic data distribution or from G. Given a data $x \sim p_{data}(x)$, the generator G and discriminator D are trained in an adversarial manner with playing a min-max game as [26]

$$\min_G \max_D \mathcal{L}_{GAN} = \mathbb{E}_{x\sim p_{data}(x)} \log\left(D\left(x\right)\right) + \mathbb{E}_{z\sim p(z)} \log\left(1 - D\left(G\left(z\right)\right)\right). \tag{3}$$

Similarly, the conditional version of GAN is more frequently used by feeding the attribute vector y into the G and D as an additional input layer. Specifically, the attribute vector y is concatenated with the prior input noise $p(z)$ in the generator, while y is taken as an input along with x into a discriminative function. As a consequence, the min-max game of conditional GAN can be denoted as [75]

$$\min_G \max_D \mathcal{L}_{CGAN} = \mathbb{E}_{x\sim p_{data}(x)} \log\left(D\left(x|y\right)\right) + \mathbb{E}_{z\sim p(z)} \log\left(1 - D\left(G\left(z|y\right)\right)\right). \tag{4}$$

3 Facial Attribute Analysis Datasets and Metrics

3.1 Facial Attribute Analysis Datasets

In this section, we introduce the overview of publicly available facial attribute datasets, including the conditions of collection, the number of subjects and image or video samples. The summary is listed in Table 1.

FaceTracer dataset is an extensive collection of real-world face images collected from the internet. There are 15,000 faces with fiducial key points marked, as well as 10 groups of attributes, totally 5,000 labels, where 7 groups facial attributes are made up of 19 attribute values, while the rest 3 groups denote the qualities of images and environment. The dataset provides the URLs of each image for considering privacy and copyright issues. Besides, FaceTracer takes 80% labeled data as training data while the remaining 20% as testing with 5-fold cross-validation.

The Labeled Faces in the Wild (LFW) dataset consists of 13,233 images of cropped, centered frontal faces that derived from the work of T. Berg et al., the Names and Faces in the News [74]. It is collected from 5,749 people using news sources online, and there are 1,680 people having two or more images. Kumar et al. [54] first collect 65 attribute labels, denoting as LFW-65 through Amazon Mechanical Turk (AMT) [1], and then expand to 73 attributes [53], denoting as LFW-73 in Table 2. Liu et al. [66] annotate LFW images with 40 face attributes and 5 fiducial key points by a professional labeling company, obtaining LFWA dataset, which is one of the most commonly used datasets in facial attribute analysis. LFWA is partitioned into half for training and the rest for testing. Specifically, there are 6,263 images for training and the remains for testing.

PubFig dataset is a large, real-world face dataset comprising 58,797 images of 200 people collected from the internet under the uncontrolled situations, which results in considerable variation in pose, light, expression, and scene, etc. It labels 73 facial attributes as many as LFW-73 but includes fewer individuals. As for the protocol, PubFig divides the development set containing 60 identity images and the evaluation set containing the rest 140 identities.

Celeb-Faces Attributes dataset (CelebA) is constructed by labeling images selected from CelebFaces [102], which is a large-scale face attributes dataset covering large pose variations and background clutter. It contains 10,177 identities, 202,599 number of face images with 5 landmark locations, as well as 40 binary attribute annotations per image. In the experiment, CelebA is partitioned into three parts: images of first 8,000 identities (with 160,000 images) for training, the images of another 1,000 identities (with 20,000 images) for validation and the rest for testing.

Berkeley Human Attributes dataset is collected from H3D [6] dataset and PASCAL VOC 2010 [108] training and validation datasets, containing 8,053 images that each centered at a full body of a person. There are wide variations in poses, viewpoints, and occlusions so that many existing methods working on front faces perform not well on the dataset. AMT is also used to provide labels for all 9 attributes by 5 independent annotators. The dataset partitions 2,003 images for training, 2,010 for validation and 4,022 for testing.

Attribute 25K dataset is collected from Facebook, which contains 24,963 people split into 8,737 training, 8,737 validation and 7,489 test examples. Since the images are with large variations in viewpoints, poses and occlusions, not every attribute can be inferred from every image. For instance, we cannot label the *wearing hat* category when the head of the person is not visible.

Ego-Humans dataset draws images from videos that track casual walkers with the OpenCV frontal face detector and facial landmark tracking in New York City throughout two months. Compared with all the datasets as mentioned above, it covers the location and weather information through clustering GPS coordinates. Moreover, nearly five million face pairs along with their same or not same labels are extracted under the constraints of temporal information and geo-location. Wang et al.

Table 1: An overview of facial attribute datasets.

Dataset	Resources	Identities / Samples	Number of attributes	Testing protocol	Access
FaceTracer [52]	Internet	15,000 / 15,000	10	Train: 80% Test: 20% Val: 5-fold cross	www.cs.columbia.edu/CAVE/databases/facetracer/
LFW [46]	Names and Faces [74]	5,749 / 13,233	65/73	Train: 50%(6,263) Test: 50%(6,970)	http://vis-www.cs.umass.edu/lfw/
LFWA [66]	LFW	5,749 / 13,233	40	Train: 50%(6,263) Test: 50%(6,970)	http://vis-www.cs.umass.edu/lfw/
PubFig [54]	Internet	200 / 58,797	73	Train: 60 identities Test: 140 identities	http://www.cs.columbia.edu/CAVE/databases/pubfig/download/
CelebA [66]	Celeb-Faces	10,177 / 202,599	40	Train: 8000 identities (160,000) Test: 1000 identities (20,000)	http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
The Berkeley Human Attributes [5]	H3D [6] PASCAL VOC 2010 [108]	- / 8,053	9	Train: 2,003 images Test: 4,022 images Val: 2,010 images	https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/shape/poselets/
Attributes 25K [118]	Facebook	24,963 / 24,963	8	Train: 8,737 identities Test: 7,489 identities Val: 8,437 identities	-
Ego-Humans[108]	Videos	- / 2,714	17	Train: 80% Test: 20%	-
University of Maryland Attribute Evaluation Dataset (UMA-ADE) [33]	Image Research	- / 2800	40	All used for test	https://www.cs.umd.edu/ emhand/research.html

Table 2: An overview of facial attributes

Dataset	Attributes								
LFW	Common		Arched Eyebrows	Attractive	Bags under eyes	Bald	Bangs	Big nose	Black hair
			Blond hair	Blurry	Brown hair	Bushy eyebrows	Chubby	Double chin	Eyeglasses
			Goatee	Gray hair	High cheekbones	Male	Mouth slightly open	Mustache	Narrow eyes
			No beard	Oval face	Pale skin	Pointy nose	Receding hairline	Rosy cheeks	Sideburns
			Smiling	Straight hair	Wavy hair	Wearing hat	Wearing lipstick	Young	
	LFW-73	LFWA/CelebA	Big lips	Heavy makeup	Wearing earrings	Wearing necklace	Wearing necktie	5'o clock shadow	
		LFW-65	Asian	Baby	Black	Child	Color photo	Curly hair	Environment
			Eyes open	Flash	Frowning	Fully visible forehead	Harsh lighting	Indian	Middle aged
			Mouth wide open	Mouth closed	No eyewear	Obstructed forehead	Posed photo	Round jaw	Round face
			Semi obscured forehead	Senior	Shiny skin	Soft lighting	Square face	Strong nose mouth lines	Sunglasses
			Teeth not visible	Teeth visible	White			(*Flushed face*	*Brown eyes*)

[108] manually annotate 2,714 images with 17 facial attributes randomly selected from these five million images. As for the testing protocol, 80% images are selected randomly for training and the rest for testing.

University of Maryland Attribute Evaluation Dataset (UMA-AED) comes from the image search taking 40 attributes as search terms and the Hyperface as the face detector [81]. UMD-AED contributes to the class-imbalanced learning in deep facial attribute estimation. It is made up of 2800 face images labeled with a subset of the 40 attributes from CelebA and LFWA. Each attribute has 50 positive and 50 negative samples, which means not every attribute is labeled in each image. Such collected way and labeled scheme make UMD-AED more presentative of real-world data than CelebA and LFWA, as well as being high-quality. Since the dataset is recently proposed and used only for a less biased evaluation, there is no relatively mature test protocol to date.

All the labels in LFW dataset with the maximum number of attributes are listed in Table 2. Different facial attribute datasets take out different subsets of these attribute annotations for deep FAE and FAE. Note that in Table 2, 'Common' denotes the attributes shared by all variant versions of LFW, totally 34 categories. LFWA and CelebA become the most commonly used FAE datasets by adding 6 attributes based on the 'Common', where the six attributes, together with the underlined *flushed face* and *brown eyes*, are added to LFW-65 to constitute the LFW-73.

3.2 Facial Attribute Analysis Metrics

3.2.1 Facial Attribute Estimation Metrics

The frequently used metrics of FAE are listed as follows.

– **The Accuracy and Error Rates**

The classification accuracy and error rates are the most commonly used measures for evaluating classification tasks, while facial attribute estimation is no exception. The accuracy rate can be defined as [86]

$$Accuracy = \left(\left(t_p + t_n\right) / \left(N_p + N_n\right)\right). \tag{5}$$

where N_p and N_n denote the numbers of positive and negative samples respectively, whereas t_p and t_p denote the numbers of true positives and true negatives [44]. Meanwhile, the error rate can be defined as

$$Error\ Rate = 1 - Accuracy. \tag{6}$$

– **The Balanced Accuracy and Error Rate**

When dealing with the class-imbalanced data, the traditional classification accuracy is not befitting due to the bias of the majority class. Hence, a balanced classification accuracy is defined as [86]

$$Balanced\ Accuracy = \frac{1}{2}\left(t_p/N_p + t_n/N_n\right). \tag{7}$$

Similarly, the balanced error rate can be defined as $Balanced\ Error\ Rate = 1 - Balanced\ Accuracy$. When dealing with the domain adaption issue [86], the balance error rate is defined as

$$Balanced\ Error\ Rate^* = \left(T^+\left(t_p/N_p\right) + T^-\left(t_n/N_n\right)\right), \tag{8}$$

where T^+ and T^- denote the target domain distribution of positive and negative examples, respectively. The superscript $*$ is used to indicate the balanced error rate in domain adaption.

– **Mean Average Precision**

As there are more than one labels in multi-label image classification, the mean Average Precision (mAP) becomes a popular metric, which computes the average of the precision−recall curve from the recall 0 to recall 1. Meanwhile, mAP is the mean of the average precision scores for a set of queries. We do not go into the details about its specific definition and formulation because of its commonality in many works [115, 80].

3.2.2 Facial Attribute Manipulation Metrics

There are two types of metrics in deep FAM: qualitative and quantitative measurements, where the former evaluates the performance of generated images through statistical surveys, and the latter gives the evaluation according to the degree of preserving the facial information after manipulation, such as identity preservation, facial landmark detection gain, and attribute prediction. We provide more detailed descriptions of the two sorts of metrics as bellow.

– **Qualitative Metrics**

Statistical survey is the most intuitive way to qualitatively evaluate the quality of generated images in most generative model construction tasks. By setting specific rules in advance, subjects vote for generated images with the appealing visual fidelity, and then, researchers draw the conclusion according to the statistical analysis of votes. For example, Choi et al. [14] quantitatively evaluate the performance of generated images in a survey format using Amazon Mechanical Turk (AMT) [1]. Given an input image, the Turkers are required to select the best generated images by instruction based on perceptual realism, quality of manipulated in attributes, and preservation of original identities. Each Turker is asked a set number of questions along with a few logical yet straightforward questions for validating human effort.

Zhang et al. [121] conduct a statistical survey as compared with prior works. Specifically, volunteers are required to choose the better result from proposed CAAE or prior works, or hard to tell via voting. Sun et al. [101] instruct volunteers to rank facial attribute manipulation approaches based on perceptual realism, quality of transferred attribute and preservation of personal features, and then calculate the average rank (between 1 and 7) of each approach. Lample et al. [56] perform a quantitative evaluation on Mechanical Turk with two different aspects of the generated images: the naturalness measuring the quality of generated images, and the accuracy measuring the degree of swapping an attribute reflected in the generation.

– **Quantitative Metrics**

Distribution difference measure helps to calculate the differences between real images and generated face images. Xiao et al. [113] achieve this goal by the Fréchet Inception Distance [42] (FID) with the means and covariance matrices of two distributions before and after editing facial attributes. Wang et al. [109] compute the Peak Signal to Noise Ratio (PSNR) to measure pixel-level differences, the Structure SIMilarity index (SSIM) and its multi-scale version MS-SSIM [110]

to estimate the structure distortion, as well as the identity distance to evaluate the high-level similarity of two face images. In light of this, face identity preservation becomes a popular evaluation for measuring the ability to preserve attribute excluding details. He et al. [41] use an Inception-ResNet [104] to train a face recognizer for evaluating the identity preservation ability with rank-1 recognition accuracy.

Facial landmark detection gain uses the accuracy gain of landmark detection before and after attribute editing to evaluate the quality of synthesized images. For example, He et al. [35] adopt a landmark detection algorithm, ERT method [50] trained on the 300-W dataset [87], to achieve this goal. During the testing, researchers partition the test sets into three components: the first one containing images with the positive attribute labels, the second containing images with the negative labels, and the last one containing the manipulated images from the first part. Then, the average normalized distance error is computed to evaluate the discrepancy of landmarks between generated images and the ground truths.

Facial attribute estimation constructs attribute prediction networks to measure the performance of FAM according to the classification accuracy. Perarnau et al. [79] first design an Anet that predicting attributes on the manipulated facial images. Once the output attribute labels of the Anet are closer to the original attribute labels, the generator can be considered to have satisfied performance. That means almost all metrics of facial attribute estimation can be used in this case.

Besides, considering that visual facial attributes produced by well-performed generative models should be correctly recognized by regression models, Larsen et al. [57] calculate the attribute similarity between the conditional attributes and generated attributes employing an attribute prediction network. Specifically, the face images are generated by retrieval from chosen attribute configurations. Then, these images are fed into a separately trained regressor network for predicting facial attributes. During the testing, faces are sampled given attributes in the test set and propagated through the attribute prediction network. As a consequence, attribute similarity scores, as well as cosine similarity and mean squared error, are all computed over the test set, where the cosine similarity is defined as the best out ten samples per attribute vector.

4 State-of-the-art Facial Attribute Estimation Methods

Generally, state-of-the-art deep FAE methods can be partitioned into two main categories: part-based methods and holistic methods. In this section, we provided an overview of these two types of methods concerning the algorithms, performance, as well as pros and cons. The summary is provided in Table 3.

4.1 Part-based Deep FAE Methods

As shown in Fig 7, deep part-based FAE methods first locate the areas where facial attributes exist by virtue of the localization mechanism. Then, features corresponding to distinct attributes on each highlighted position can be extracted and further be predicted with multiple attribute classifiers. Note that the key of part-based methods lies in the localization mechanism. In light of this, deep part-based methods can be further divided into two sub-groups: separate auxiliary localization and end-to-end localization. In this section, we will provide more details about the two types of localization schemes, respectively.

4.1.1 Separate Auxiliary Localization

Since facial attributes describe subtle details of face representations based on human vision, locating the positions of facial attributes enforces subsequent feature extractors and attribute classifiers to focus more on attribute-relevant regions. The most intuitive way is taking existing face part detectors as auxiliaries.

Poselet [6,5] is a valid part detector which describes a part of the human pose under a given viewpoint. Since these parts include evidences from different parts of the body at different scales, complementary information can be learned to benefit attribute prediction. Typically, given a whole person image, poselet detector [118] is first used to decompose images into several image patches, named as poselets, under various viewpoints and poses. Then, a PANDA network that trains a CNN per poselet and concatenates features from all these poselets, as well as the whole image, is designed to yield the final part-based deep representations. Finally, PANDA branches out multiple binary classifiers where each recognizes an attribute through the binary classification. Based on PANDA, Gkioxari et al. [24] introduce a deep version of poselets and build a feature map pyramid whose each level computes a score for the corresponding attribute prediction.

However, poselet detector discovers the coarse body parts and cannot explore local details of face images. Considering the probability of an attribute appearing in a face image is not uniformed in the spatial domain, Kalayeh et al. [49] propose to employ the semantic segmentation as a separate auxiliary localization scheme

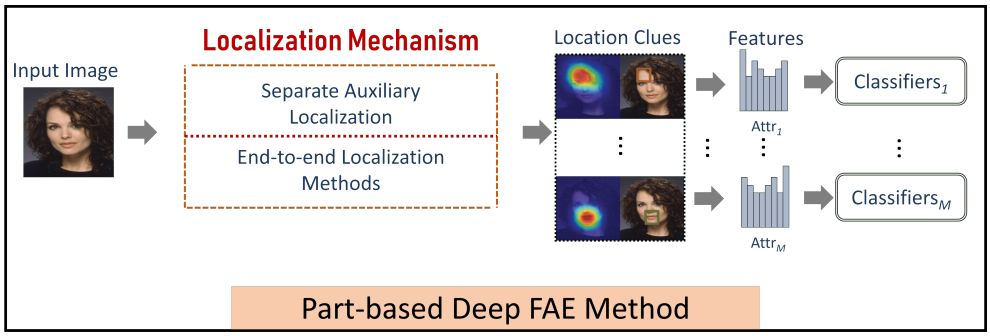


Fig. 7: The illustration of deep part-based FAE methods.

to build a prediction model. They exploit the localization cues obtained by the semantic segmentation to guide the attention of attribute prediction to the naturally occurring areas of different attributes. Looking at the details, as regards for the model construction, first, a semantic segmentation network is designed in the form of an encoder-decoder and trained over Helen face dataset [58]. During this process, the semantic face parsing [96] is taken as an additional task to learn detailed pixel-level localization information. Once the localization cues from the semantic segmentation network are discovered, the semantic segmentation-based pooling (SSP) and gating (SSG) mechanisms are presented to integrate the location information into the attribute estimation. SSP decomposes the activations of the last convolution layer into distinct semantic regions and then aggregates those regions that only reside in the same area. In the meanwhile, SSG gates the output activations between the convolution layers and the batch normalization (BN) operations to control the activations of neurons from different semantic regions.

Rather than taking semantic segmentation as the auxiliary locating task, Mahbub et al. [72] consider a more straightforward way by segmenting faces into several image patches directly according to the key point marks. Then, these segments are fed into a set of facial segment networks to extract corresponding feature representations and learn prediction scores, along with the whole face image flowing into a full-face network. A global predictor network fuses the features from these segments, and two committee machines merge their scores for the final prediction.

In contrast to the above two methods that look for location clues of attributes directly, He et al. [37] harness synthesized abstraction facial images that contain local facial parts and texture information, to achieve the same goal indirectly. A generative adversarial network is used to generate such facial abstraction images before inputting them into a dual-path facial attribute recognition network, along with the real original images. The dual-path network propagates the feature maps from the abstraction sub-network to the real original image sub-network and concatenates the two types of features for the final prediction. Despite the abundant location and textual information of generated facial abstractions images, the quality of these images may be a significant performance hit, as some subtle attribute information may lose in the process of image abstraction.

Note that all the separated auxiliary localization based deep FAE methods share a common drawback: relying on too much accurate facial landmark localization, face detection, facial semantic segmentation and parsing, as well as facial partition according to specific criteria in the training process. Once these localization strategies are imprecise or landmark annotations are unavailable, the performance of FAE would deteriorate.

4.1.2 End-to-end Localization

Compared with the separate auxiliary localization methods that locate attribute regions and make the attribute prediction separately and independently, end-to-end localization methods jointly exploit the localizations where facial attributes show up and predict their presences in a unified framework.

Liu et al. [66] first propose a cascaded deep learning framework for jointly face localization and attribute prediction. Specifically, the cascaded CNN is made up of an LNet and an ANet, where the LNet locates the entire face region and the ANet extracts high-level face representation from the located area. When it comes to the details of the LNet, it is firstly pre-trained by classifying massive general object categories to ensure the excellent generalization capability, and then fine-tuned using the image-level attribute tags of training images to learn useful features for face localization in a weakly supervised manner. Note that the critical difference between the LNet and the separated auxiliary localization methods is that the LNet has no more need

to face bounding boxes or landmark annotations. As for the ANet, first, it is pre-trained by classifying massive face identities to cope with complex variations in the unconstrained face images and then fine-tuned by attribute estimation to extract discriminative face representations. Furthermore, rather than extracting features patch-by-patch, ANet evaluates images with a fast feed-forward scheme, in which the one-pass feed-forward operation with locally shared filters and an interweaved operation is leveraged to learn the discriminative feature representations. Finally, SVMs are trained over these features to estimate attribute values per attribute, while the terminal prediction is made by averaging all values to deal with the small misalignment of face localization. The cascaded LNet and ANet framework shows the benefit of pre-training with massive object categories and massive identities in improving the learning of feature representations. With such customized pre-training schemes and cascaded architecture for jointly face localization and attribute prediction, the method shows outstanding robustness to the background and face variations.

However, LNet only discovers coarse entire face regions, more local attribute details are not fully explored. For overcoming this, Ding et al. [16] propose a cascade network to jointly learn to locate facial attribute-relevant regions and perform attribute classification. Specifically, they firstly design a face region localization network (FRL) that builds a branch for each attribute to automatically detect a corresponding relevant region. Then, the followed parts and whole (PaW) attribute classification network selectively leverage information from all the attribute-relevant regions for the final estimation. Moreover, during the attribute classification, Ding et al. define two full connected layers: the region switch layer (RSL) and the attribute relation layer (ARL). The former selects the relevant sub-network for predicting attributes while the latter models attribute relationships. In a word, the cascaded FRL and PaW model not only discover semantic attribute regions but also explore rich relationships among facial attributes. Besides, since the model automatically detects face regions, the alignment of the face image is not required, so that it even achieves outstanding performance over unaligned datasets.

Note that FRL-PaW method learns a location for each attribute, which makes the training process redundant and time-consuming as several attributes exist in the same area. However, to the best of our knowledge, there is no specific solution for tackling this issue so far.

Current FAE methods are more inclined to take the whole face images as inputs without paying attention to where attributes are and design networks to model correlations among attributes. This is precisely what the holistic deep FAE methods do, and we will provide a detailed introduction of them in the following section. Before that, we summarize the part-based deep FAE methods as follows.

Part-based deep FAE methods first locate the positions where facial attributes appear. Two strategies can be adopted: separate auxiliary localization and end-to-end localization, where the former leverages existing part detectors or auxiliary localization-related algorithms, while the latter jointly exploits the localizations in which facial attributes exist and predicts their presences. Compared with the separate auxiliary localization methods operating separately and independently, end-to-end localization methods locate and predict in a unified framework. Once the localization clues are obtained, features corresponding to certain areas related to specific attributes can be extracted and further fed into attribute classifiers to be estimated.

4.2 Holistic Deep FAE Methods

In this section, we introduce holistic deep FAE methods, and the schematic diagram of model construction is provided in Fig. 8.

In contrast to part-based approaches detecting and utilizing facial components, holistic deep FAE methods focus more on exploring the attribute relationships and extracting features from entire face images with no need for facial parts. The key of modeling attribute relationships is learning common features at low-level shared layers and exploring attribute-specific features at high-level separated layers, where each separated layer corresponds to an attribute group. In general, these attribute groups are obtained by manual according to semantics or attribute locations. In this way, through assigning different shared layers and attribute-specific layers, complementary information among multiple attributes can be discovered, so that more discriminative features can be learned for the following attribute classifiers.

To sum up, there are two crucial issues that holistic deep FAE methods need to deal with when designing network architectures:

(1) how to properly assign shared information and attribute-specific information at different layers of networks?

(2) how to explore relationships among distinct attributes for learning more discriminative features?

Existing methods have made many efforts to solve the two problems, and we provide a brief review of these methods as below.

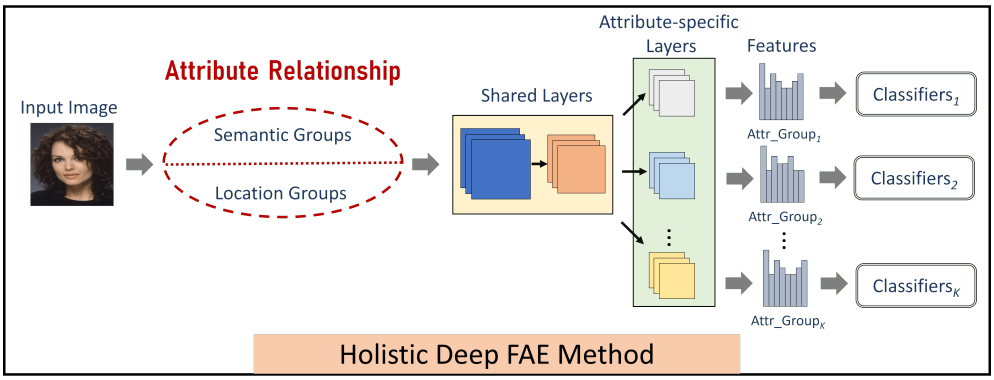


Fig. 8: The illustration of deep holistic FAE methods.

To our best knowledge, the earliest method using the holistic multi-task framework for deep FAE is MOON [86], mixed objective optimization network, which learns multiple attribute labels simultaneously via a single DCNN. MOON takes facial attribute estimation as a regression problem for the first time and adopts a 16-layer VGG network as the primary network configuration, in which abstract high-level features are shared before the last FC layer. Consequently, prediction scores of multiple attributes can be calculated with the MSE loss for reducing the regression error. Similarly, Zhong et al. [123] take the place of high-level CNN features in MOON with mid-level ones to identify the best representations for each attribute.

Compared with splitting networks at the last FC layer, Hand et al. [34] present a multi-task deep CNN (MCNN), which branches out to multiple groups at the mid-level convolutional layers, to model the correlations among facial attributes. Specifically, based on the assumption that many attributes are strongly correlated, MCNN divides all the 40 attributes into 9 groups according to semantic, i.e., gender, nose, mouth, eyes, face, around head, facial hair, cheeks, and fat. For example, *big nose* and *pointy nose* are grouped into the 'nose' category, while *big lips*, *lipstick*, *mouth slightly open* and *smiling* are clustered into the 'mouth' category. Therefore, each group consists of similar attributes and learns high-level features independently. In the first two convolutional layers of MCNN, features are shared by all attributes. After this, MCNN branches out several forks corresponding to conspicuous attribute groups, which means each attribute group possesses a fork. At the end of the network, an FC layer is added to create a two-layer auxiliary network (AUX) to facilitate attributes relationships. AUX receives the scores from the trained MCNN and products the ultimate prediction results. To sum up, MCNN-AUX models the attribute relationships in three ways: (1) sharing the lowest layers for all attributes; (2) assigning the higher layers for spatially-related attributes; (3) discovering score-level relationships with the AUX network.

However, there exists a limitation in MCNN that shared information at low-level layers may vanish after network splitting. One solution to tackle this restriction is jointly learning shared and attribute-specific features at the same level, rather than in order of precedence.

Cao et al. [8] design a partially shared structure based on MCNN, i.e., PS-MCNN. It divides all 40 attributes into 4 groups according to attribute positions by manual, i.e., upper group, middle group, lower group, and whole image group. Note that such a manual grouping strategy can be regarded as prior information based on human knowledge. The partially shared structure connects four attribute-specific networks (TSNets) corresponding to four different groups of attributes and one shared network (SNet) sharing features among all the attributes. Specifically, each TSNet learns features for a specific group of attributes, while SNet shares informative features with each task. As for the connection mode between the two types of sub-networks, each layer of SNet receives additional inputs from the previous layers of TSNet. Then, features from SNet are fed into the next layers of shared and attribute-specific networks. In a word, PS-MCNN learns features at a certain layer based on both task-specific features and shared features. Besides, shared features at a specific layer are closely related to the features of all its previous layers, leading to informatively shared representations.

Apart from the attribute correlations, Han et al. [32] introduce the concept of attribute heterogeneity. They point out that individual attributes could be heterogeneous concerning data type and scale, and semantic meaning. In terms of data type and scale, attributes can be grouped into ordinal vs. nominal attributes. For instance, if attributes, like *age* and *hair length*, are ordinal, attributes like gender and race are nominal. Note

that the difference between ordinal and nominal attributes is that ordinal attributes have an explicit ordering of their variables, while nominal attributes usually have two or more classes and there is no intrinsic ordering among the categories.

Similarly, in terms of semantic meaning, attributes such as *age*, *gender*, and *race* are used to describe the characteristics of the whole face, and the *pointy nose* and *big lips* are mainly used to describe the local characteristics of facial components. Therefore, these two categories of attributes are heterogeneous and can be grouped into holistic vs. local attributes for the prediction of different parts of a face image.

Therefore, taking both the attribute correlation and heterogeneity into consideration, Han et al. design a deep multi-task learning (DMTL) CNN that learns shared features of all attributes and category-specific features of heterogeneous attributes. The shared feature learning naturally exploits the relationship among attributes to achieve robust and discriminative feature representations, while the category-specific feature learning aims at fine-tuning the shared features towards the optimal estimation of each heterogeneous attribute category.

Note that existing multi-task learning methods make no distinction between all low-level and mid-level features for different attributes, which is unseemly as features at different levels of networks may have different relationships. Besides, the above methods share features across tasks and split layers that encode attribute-specific features by hand-designed network architectures. Such a manual exploration in the space of possible multi-task deep architectures is tedious and error-prone, as possible spaces may be combinatorially large.

Therefore, rather than constructing networks by manual, Lu et al. [67] present a case that automatically designs compact multi-task deep learning architectures with no need of discovering the possible multi-task architectures artificially. The proposed network learns shared features in a fully adaptive way. Specifically, it starts with a thin multi-layer network (VGG16) and dynamically widens in a greedy manner during training, so that both task correlations and complexity of the model can be taken into account, enabling task grouping decisions at each layer of the network. The initialization of the thin network is based on a simultaneous orthogonal matching pursuit (SMOP) [107], resulting in faster convergence and higher accuracy.

The core of the fully adaptive feature sharing algorithm is incrementally widening the current design in a layerwise fashion. Therefore, a top-down layer-wise model widening strategy is adopted. During the training process, the network decides with whom each task shares features in each layer, which leads to several branches in this layer. Thus, the whole process is fulfilled in a multi-round branching manner. Finally, the number of branches at the last layer of the model is equal to that of attribute categories to be predicted. This method allows us to estimate multiple facial attributes in a dynamic branching procedure through its self-constructed architecture, as well as the fully adaptive feature sharing strategy.

5 State-of-the-art Facial Attribute Manipulation Methods

As we illustrated in Fig. 6, in terms of the model construction, state-of-the-art deep FAM methods can be divided into two categories: model-based methods and extra condition-based methods. In this section, we provide an overview of these two types of methods concerning the algorithms, network architectures, as well as pros and cons. The summary is provided in Table 4.

5.1 Model-based Deep FAM Methods

Model-based methods map an image in the source domain to the target domain, and then distinguishes the generated target distribution with the real target distribution under the constraint of an adversarial loss. Such model-based methods are greatly task-specific, resulting in more photo-realistic facial attribute images.

Typically, Li et al. [61] first propose a DIAT model following the standard paradigm of model-based methods. DIAT takes unedited images as inputs to generate target facial images, which not only possess target attributes with an adversarial loss but also keep the same or similar identity to the input images with an identity loss. Zhu et al. [125] add an inverse mapping from the target domain to the source domain based on DIAT and propose a CycleGAN, where the two mappings are coupled with a cycle consistency loss. This process is based on the intuition that if we translate from one domain to the other and back again, we should arrive where we start. Based on CycleGAN, Liu et al. [65] propose a UNIT model that maps the pair of corresponding images in source and target domains to a same latent representation in a shared latent space. Each branch from one of domains to the latent space implements a analogous CycleGAN operation.

However, all the above methods operate on the whole face image directly. That means when a certain attribute is edited, the rest of other relevant attributes may be changed uncontrollably in the meanwhile.

Therefore, in order to modify the attribute-specific face areas and keep the other parts unchanged, Shen

Table 3: The State of Art Deep Facial Attribute Estimation Approaches.

Categories	Approaches	Algorithms	Network Architectures	Datasets	Metrics and Performance
Part-based Methods	PANDA [118](CVPR2014)	Using Part-based Pose Aligned Networks for learning features related to poses and Linear SVM classifiers for attribute estimation	PANDA	Berkeley Human Attributes Dataset Attributes 25K Dataset LFW-gender	mean Average Precision (mAP) The Berkeley Human Attributes Dataset (78.98%) Attribute 25K Datasets(70.74%) LFW-gender (99.54%)
	Gkioxari et al. [24](ICCV2015)	Using a deep version of poselets and capturing parts of the human body for tasks of action and attribute classification	A 5-layer CNN feature pyramid and a pyramid of part scores	Berkeley Human Attributes Dataset	mAP(89.5%)
	LNet+ANet [66](ICCV2015)	Cascading LNet CNN for localization and ANet for feature extraction	LNet+ANet	CelebA LFWA	Accuracy CelebA(87%) LFWA(84%)
	Off-the-shelf CNN [122](ICB2016)	Training off-the-shelf architectures for face recognition to construct facial representations	Off-the-shelf	CelebA LFWA	Accuracy CelebA(86.6%) LFWA(84.7%)
	Singh et al. [95](ECCV2016)	Using Spatial Transformer Network (STN) and Ranker Network (RN) to jointly learn features, localization and ranker of attributes	STN and RN	LFW-10attr	Attribute ranking accuracy (86.91%)
	SSP+SSG [49](CVPR2017)	Using semantic segmentation guiding the attention of the attribute prediction to the regions where different attributes naturally show up	Semantic Segmentation-based Pooling(SSP) Semantic Segmentation-based Gating(SSG)	CelebA	Error Rate(8.20%) Average Precision(81.45%) Balanced Accuracy(88.24%)
	FRL-PaW [16](AAAI2018)	Simultaneously learning to localize face regions specific to attributes and performs attribute classification without alignment in a cascade network	Facial region localization (FRL) network Parts and Whole (PaW) classification network	Unaligned CelebA	Accuracy(91.23%)
	SPLITFACE [72](arxiv2018)	Using facial segmentation for attribute detection in partially occluded faces	Segmentwise, Partial, Localized Inference in Training Facial Attribute Classification Ensembles (SPLITFACE) Network	CelebA	Accuracy(90.61%)
	FMTNet [127](PR2018)	Constructing three sub-networks for attribute transfer learning	the Face detection Network (FNet) the Multi-label learning Network (MNet) the Transfer learning Network (TNet)	CelebA LFWA	Accuracy CelebA(91.66%) LFWA(84.34%)
	He et al. [37](IJCAI2018)	Generating abstraction images by GAN as complementary features and used for facial parts localization	GAN and a dual-path facial attribute recognition network	CelebA LFWA	Accuracy CelebA(91.81%) LFWA(85.2%)
	AFFAIR [60](TIP2018)	Learning a hierarchy of spatial transformations for facial attribute prediction with no landmark	lAndmark Free Face AttrIbute pRediction (AFFAIR) Network	CelebA LFWA MTFL	mean AP/Accuracy CelebA(79.63%/91.45%) LFWA(83.01%/86.13%) MTFL(-/86.55%)
Holistic Methods	Wang et al. [108](CVPR2016)	Employing a Siamese structure and embedding location as well as weather contextual information for learning feature representation	Siamese	CelebA LFWA Ego-Humans Dataset	Accuracy CelebA(88%) LFWA(87%) Ego-Humans Dataset(87%)
	MOON [86](ICCV2016)	Treating attribute classification as a regression task and solving domain adaptive problem	Mixed-Objective Optimization Network (MOON, VGG16-based)	CelebA	Error Rate CelebA(9.06%) A balance error rate CelebAB(13.67%)
	LMLE [44](CVPR2016)	Using a Large Margin Local Embedding (LMLE) Method for large-scale imbalanced classification tasks of binary facical attributes	VGG-6 Quintuplet CNN	CelebA	Balanced Accuracy(84.25%)
	Zhong et al. [123](ICIP2016)	Studying the effect of mid-level CNN features for attribute prediction	FaceNet NN.1 [90]	CelebA LFWA	Accuracy CelebA (89.8%) LFWA (85.9%)
	CRL [18](ICCV2017)	Combining batch-wise incremental hard mining for class-imbalance with the Class Rectification Loss (CRL) regularizing algorithm for attribute classification	5-layer DeepID2 [102] CNN	CelebA	Balance Accuracy(86%)
	AFFACT [29](IJCB2017)	Introducing the Alignment-Free Facial Attribute Classification Technique(AFFECT) such data augmentation technique for attribute classification without alignment	AFFACT Network (ResNet based)	CelebA	Error Rate(8.03%)
	MCNN+AUX [34](AAAI2017)	Considering attribute relationships and constructing a Multi-task deep CNN(MCNN) with an Auxiliary Network(AUX) for performance improvement	MCNN+AUX	CelebA LFWA	Accuracy CelebA (91.22%) LFWA (86.31%)
	DMTL [32](TPAMI2017)	Introducing Deep multi-task feature learning (DMTL) for joint estimation of multiple heterogeneous attributes	DMTL(AlexNet based)	CelebA LFWA	Accuracy CelebA (93%) LFWA (86%)
	Lu et al.[67](CVPR2017)	Automatically designing compact multi-task deep learning architectures starting with a thin multi-layer network and dynamically widening in a greedy manner	Automatic top-down layer wise widening	CelebA	Accuracy(91.02%) Top-10 Recall(71.38%)
	AttCNN [33](AAAI2018)	Selectively learning with domain adaptive batch re-sample methods for multi-label attribute prediction	AttCNN Network	CelebA LFWA UMD-AED	Accuracy Balanced Domain CelebA(85.05%) LFWA(73.03%) UMD-AED(71.11%)
	R-Codean [91](PRLetters2018)	Incorporating a Cosine similarity based loss function into the Euclidean distance for constructing an R-Codean autoencoder	Residual Codean Autoencoder	CelebA LFWA	Accuracy CelebA(90.14%) LFWA(84.90%)
	PS-MCNN [8](CVPR02018)	Considering the identity information and attribute relationships simultaneously and constructing a Partially Shared Multi-task Convolutional Neural Network (PS-MCNN)	PS-MCNN	CelebA LFWA	Error rates CelebA (7.02%) LFWA (12.64%)

et al. [92] present learning residual images so that face attributes can be manipulated efficiently with modest pixel modification over the attribute-specific regions. The residual images are defined as the discrepancy between images before and after the attribute manipulation. A ResGAN, consisting of two image transformation networks and a discriminative network, is designed to learn the residual representations of desired attributes. Specifically, the transformation networks take responsibility for the attribute manipulation and its dual operation, whereas the discriminative network distinguishes the generated images from real images. Two image transformation networks, which are denoted as G_0 and G_1 respectively, firstly take two images with opposite attributes as inputs in turn, and then perform the inverse attribute manipulation operation for outputting

residual images. After this, the obtained residual images are added to the original input images, yielding the final outputs with manipulated attributes. In the end, all these images, i.e., two original input images and two images from the transformation networks, are both taken as inputs of the discriminative network, which classifies these images into three categories, i.e., images generated from the two transformation networks, original images with positive attribute labels and original images with negative attribute labels.

Moreover, drawing on the dual learning applied in machine translation [35], a given image with a negative attribute label flows into G_0 for synthesizing the desired attribute, then the obtained image is fed to G_1 for removing the synthetic attribute. In this cycle, G_0 is the primal task, while G_1 is regarded as the dual task of G_0. Then, the yielded image of G_1 is expected to have the same attribute label with the original given image. The experiments demonstrate that the dual learning process is beneficial for the generation of high-quality images, and residual images based facial attribute manipulation can successfully enforce the process of synthesis to pay more attention to local areas where attributes show up, especially for those local attributes.

Since no extra conditional constraints are requested, model-based methods can only edit an attribute during a training process with a set of corresponding model parameters, and the whole manipulation is only supervised through discriminating real or generated images with the adversarial loss. That means when multiple attributes need to be changed, multiple training processes are inevitable, causing the time-consuming problem.

In contrast, manipulating facial attributes with extra conditions is a more commonly used way, as multiple attributes can be edited through a single training process by controlling multiple attributes. Hence, extra condition-based methods attract more attention of researchers, where extra attribute vectors and reference exemplars are taken as input conditions. Specifically, attribute vectors can be concatenated with the latent image codes for controlling facial attributes, while reference exemplars exchange specific attributes with the to-be-manipulated images in the image-to-image translation framework. More details about the extra condition-based deep FAM methods are introduced below.

5.2 Extra Condition-based Deep FAM Methods

Deep FAM methods conditioned on extra attribute vectors alter desired attributes with given conditional attribute vectors, such as one-hot vectors indicating the presence of corresponding facial attributes. During the training, the conditional vectors are concatenated with the to-be-manipulated images in latent spaces. Moreover, as we all know, conditional generative frameworks dominate model constructions of deep FAM algorithms. In this case, various efforts are made to edit facial attributes over the autoencoder (AE), VAE, and GAN.

Zhang et al. [121] propose a conditional adversarial autoencoder (CAAE) for the age progression and regression purposes. CAAE first maps a face image to a latent vector through an encoder. Then, the obtained latent vector concatenated with an age label vector is fed into a generator for learning a face manifold. The age label condition controls to alter the age, while the latent vector ensures the personalized face features are preserved. Yan et al. [114] introduce a conditional variational autoencoder (CVAE) to generate images from visual attributes. CVAE disentangles an image into the foreground and the background parts, where each part is combined with the defined attribute vector, respectively. As a consequence, the generation quality of complex images can be significantly improved, since much more attention is paid to the pivotal foreground areas. Perarnau et al. [79] propose an invertible conditional GAN (IcGAN) to edit multiple facial attributes with determined specific representations of generated images. Given an input image, IcGAN firstly learns a representation containing a latent variable and a conditional vector via an encoder. Then, IcGAN modifies such the latent variable and conditional vector to regenerate the original input image through the conditional GAN [75]. In this way, through changing the encoded conditional vector, IcGAN can achieve arbitrary attribute manipulation.

Apart from the autoencoder, VAE, as well as GAN and their variations, Larsen et al. [57] combine the VAE and GAN into an unsupervised generative model, termed as, VAE/GAN. In this model, GAN discriminator learns feature representation taken as the basis of the VAE reconstruction objective, which means the VAE decoder and the GAN generator are collapsed into one by sharing parameters and training jointly. Hence, this model contains three parts: the encoder, the decoder, and the discriminator. By concatenating attribute vectors with feature representations from such three components, VAE/GAN performs better than either of plain VAEs and GANs.

Recently, taking the multiple attribute manipulation as a domain transfer task, Choi et al. [14] propose a StarGAN to learn mappings among multiple domains with only a single generator and a discriminator training from all domains, where each domain corresponds to an attribute and the domain information can be denoted by one-hot vectors. Specifically, first, the discriminator distinguishes the real and fake images, as well as

Table 4: The State of Art Facial Attribute Manipulation Approaches.

Categories	Approaches	Algorithms	Network Architectures	Datasets
Model-based	DIAT [61] (arxiv1610)	Transferring input images to each reference attribute label while keeping the same or similar identity for Identity-Aware Transfer (DIAT) of facial attributes	GAN	CelebA
	InfoGAN [12] (NIPS2016)	Maximizing mutual information for interpretable representations and discovering visual concepts of facial attributes	GAN	CelebA
	UNIT [65] (NIPS2017)	Proposing an UNsupervised Image-to-Image Translation (UNIT) framework under a shared-latent assumption	GAN+VAE	CelebA
	Residual Image [92] (CVPR2017)	Learning residual images to avoid entire face operation with redundant irrelevant information	GAN	CelebA
	Wang et al. [109] (WACV2018)	Combining a perceptual content loss and two adversarial losses to guarantee the global consistency for producing more realistic images	GAN	CelebA LFW
	SG-GAN [117] (ACMMM1805)	Constructing a sparsely grouped generative adversarial networks (SG-GAN) in the sparsely grouped datasets where most training data are mixed and a few are labelled	GAN	CelebA
Extra Condition-based	*CONDITIONED ON ATTRIBUTE VECTORS*			
	VAE/GAN [57] (ICML2016)	Using learned feature representations in the GAN discriminator as basis for the VAE reconstruction objective	GAN+VAE	LFW
	CVAE [114] (ECCV2016)	Learning a layered foreground-background generative conditional variational auto-encoder for complex images	VAE	LFW
	IcGAN [79] (arxiv1611)	Combining an encoder with a cGAN for obtaining Invertible cGAN (IcGAN)	GAN+VAE	CelebA
	Fader Network [56] (NIPS2017)	Disentangling the salient information of face images and the values of attributes directly in the latent space for modifying facial attributes continuously	AE	CelebA
	CAAE [121] (CVPR2017)	Learning a face manifold for smooth age progression and regression simultaneously in a conditional adversarial autoencoder (CAAE)	AE	FGNET
	cCycleGAN [68] (ECCV2018)	Extending the cycleGAN [125]conditioned on facial attributes with the cycle consistency loss	GAN	CelebA
	StarGAN [14] (CVPR2018)	Constructing a StarGAN for multiple domain image-to-image translations	GAN	CelebA
	CRGAN [59] (Springer2018)	Introducing recycle reconstruction loss to maintain personal facial identity and directly learning facial transformation with attribute annotations	GAN	CelebA
	SaGAN [116] (ECCV2018)	Introducing a spatial attention mechanism for only modifying the attribute-specific region and keep the rest unchanged	GAN	CelebA
	CONDITIONED ON REFERENCE EXEMPLARS			
	Gene-GAN [124] (arxiv1705)	Recombing the latent representation information of two paired attribute images for swapping specific attributes	GAN	CelebA
	ELEGANT [113] (ECCV2018)	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	GAN+VAE	CelebA
	EGSC-IT [71] (ICLR2019)	Constructing an exemplar guided semantically consistent image-to-image translation (EGSC-IT) network to control the translation process under exemplar images in the target domain.	GAN+VAE	CelebA

classifies the real images to its corresponding domain. Then, the generator is trained to translate an input image into an output image conditioned on a target domain label vector, which is generated randomly. As a result, the generator is capable of translating the input image flexibly. In a word, StarGAN takes the domain labels as extra supervision conditions, which makes it possible to incorporate multiple datasets containing different types of labels simultaneously.

As we can see, all the above methods achieve to edit multiple facial attributes simultaneously by changing multiple values of attribute vectors. However, none of them can implement such operation continuously.

In light of this, Lample et al. [56] present a Fader network using continuous attribute values to modify attributes through sliding knobs, like faders on a mixing console. For example, one can gradually change the values of *gender* to control the transition process from *man* to *woman*. Fader network is composed of three ingredients: an encoder, a decoder, and a discriminator. With an image-attribute pair as the input, Fader network first maps the image to the latent representation by its encoder and predict the attribute vector by its discriminator. Then, the decoder reconstructs the image through the learned latent representation and attribute vector. At the testing time, the discriminator is discarded, and different images with various attributes can be generated with different attribute values.

Note that all the above methods edit attributes over the whole face images. Such strategy brings an apparent drawback: attribute-irrelevant details might be changed in the meanwhile. Hence, how to keep other areas that are outside of the specific attribute-relevant regions unchanged is still a challenge to face.

In order to tackle this, Zhang et al. [116] introduce the spatial attention mechanism into GANs to locate attribute-relevant areas for manipulating facial attributes more precisely. The proposed GAN with spatial attention, dubbed SaGAN, follows the standard adversarial learning paradigm, where a generator and a discriminator play a min-max game. Meanwhile, extra attributes vectors are used for editing specific attributes. To keep the attribute-irrelevant regions unchanged, the generator consists of an attribute manipulation network (AMN) and a spatial attention network (SAN). Given a face image, SAN learns a spatial attention mask where attribute-relevant regions have non-zero attention values. In this way, the region which the desired attribute show up can be located. After this, AMN takes the face image and the attribute vector as inputs, yielding an image with the manipulated attribute in the specific region located by SAN.

Rather than taking the attribute vectors as extra conditions, deep FAM methods conditioned on reference exemplars consider to exchange specific attributes with the to-be-manipulated images in the image-to-image translation manner. Note that these reference images do not need to have the same identity with the original to-be-manipulate images, and all the generated attributes are present in the real world. In this way, more specific details that appear in the reference images can be explored to generate more realistic images other than altering attribute vectors manually.

Typically, Zhou et al. [124] first design an GeneGAN to achieve the basic reference exemplar-based facial attribute manipulation, where an image is encoded into two complement codes, i.e., attribute-specific codes and attribute-irrelevant codes. By exchanging the attribute-specific codes between the reference exemplars and to-be-manipulated images, desired attributes can be transferred from one image to another.

Considering that GeneGAN only transfers one attributes, Xiao et al. [113] construct an ELEGANT model to exchange latent encodings with GAN for transferring multiple face attributes by exemplars. Specifically, since all the attributes are encoded in the latent space in a disentangled manner, one can exchange the specific part of encodings and manipulating several attributes simultaneously. Besides, the residual image learning and the multi-scale discriminators for adversarial training enable the proposed model to train on higher resolution images and generate high-quality images with more delicate details and fewer artifacts, respectively. At the beginning of training, ELEGANT receives two sets of training images as inputs, i.e., a positive set and a negative set, which are not necessary to be paired. Second, an encoder is utilized to obtain the latent encodings of both positive and negative images. After this, if the i-th attribute is required to be transferred, the only thing that needs to do is to exchange the i-th element in the latent encodings of positive and negative images. Once the encoding part is finished, a reasonable structure should be designed for deciphering the latent encodings into images. ELEGANT recombines the latent encodings and employs a decoder to do this job, along with the encoder, together playing a role as the image generator. At last, the multi-scale discriminators are utilized, consisting of two discriminators with identical network structure but operating at different scales, to obtain the final manipulated facial attribute images.

6 Additional Related Issues

6.1 Class-imbalanced Learning in Facial Attribute Analysis

Class-imbalanced data refer that, in a dataset, some of the classes have a much higher number of samples than others, corresponding to the majority class and minority class [31], respectively. For example, the largest imbalance ratio between the minority and majority attributes in CelebA dataset is 1:43. In the most real-world applications, the frequently used schemes for tackling this problem are data re-sampling or cost-sensitive learning. In the following, we list several strategies used in the field of deep facial attribute analysis.

MOON [86] weights the back propagation error according to a given balanced target distribution in a cost-sensitive way. Specifically, MOON takes the class-imbalanced problem as the domain adaption task, which utilizes the balanced target domain distribution to instruct the imbalanced source domain. Observing that MOON does not take the label imbalance over each batch into account, AttCNN [33] applies a domain adaptive re-sampling at the batch level via the proposed selective learning algorithm. Consequently, one can correct the bias of data in each batch according to the desired distribution.

Rather than handling class-imbalanced data in the form of domain adaption, Huang et al. [44] formulate a quintuple sampling method with the associated triple-header loss, called large margin local embedding (LMLE) from the perspective of re-sampling. LMLE enforces the preservation of locality cross clusters and discrimination between classes. Then, a fast cluster-wise kNN is executed followed by a local large margin decision. In this way, LMLE learns embedded features, which are discriminative enough without any possible local class imbalance.

However, Dong et al. [18] point out that LMLE has several weaknesses, including the separate process of feature extraction and classification, quintuplet computing costs, and offline clustering of data. In light of this, Dong et al. [18] propose an end-to-end method to ameliorate these drawbacks. In detail, they exploit a batch-wise incremental hard mining on minority attribute classes, and in the meantime, formulate a class rectification loss (CRL) based on the mined minority examples. As for the hard mining strategy, to begin with, the profiles of minority hard-positives and hard negatives are provided, and then for a minority class of a specific attribute, K hard-positives as the bottom-K scored on the minority class are selected, as well as K hard-negatives as the top-K scores, given the predefined profiles and model. Such process is executed at the batch level and incrementally over subsequent batches, i.e., batch-wise incremental hard mining.

Based on LMLE, Huang et al. [45] present a cluster-based large margin local embedding (CLMLE), the rectified version of LMLE. CLMLE aims to learn more discriminative deep representations with a CLMLE loss via the preservation of inter-cluster margin both within and between classes. Different from LMLE enforcing Euclidean distance on a hypersphere manifold, CLMLE designs angular margins enforced between the involved cluster distributions and adopt spherical k-means for obtaining K clusters with the same size. CLMLE achieves better performance than CRL and LMLE methods.

6.2 Relative Attribute Ranking in Facial Attribute Analysis

Relative attribute learning aims at formulating functions to rank the relative strength of attributes [11], which can be widely applied in objection detection [21], fine-grained visual comparison [93], and facial attribute estimation [60]. The general insight in this line of work is learning global image representations in a unified framework [55, 77], or localizing part-based representations through pre-trained part detectors [5, 89, 118] . However, the former ignores the localizations of attributes, while the latter ignores the correlations among attributes. That means both the two might collapse the performance of relative attribute ranking.

Xiao et al. [112] first propose to automatically discover the spatial extent of relevant attributes by establishing a set of visual chains indicating the local and transitive connections. In this way, the localizations of attributes can be learned automatically in an end-to-end way. Although no pre-trained detectors are used, the optimization pipeline still contains several independent modules, resulting in the suboptimal solution.

Singh et al. [95] construct an end-to-end deep CNN for learning the features, localizations, and ranks of facial attributes simultaneously for tackling this issue by taking the weakly-supervised pairwise images as inputs. Specifically, given pairs of training images ordered according to the relative strength of an attribute, two Siamese networks receive these images, where each takes one of a pair as input and leads to a single branch. Each branch contains two components: one is the spatial transformer network (STN), which generates image transformation parameters for localizing the most relevant regions, and another is the ranker network (RN), which outputs the predicted attribute scores for the images. The qualitative results on LFW-10 dataset show good performance on attribute region localization and ranking accuracy.

To model the pair-wise relationships between images for multiple attributes, Meng et al. [73] construct a graph model, where each node represents an image and edges model the relationships between images and attributes, as well as between images and images. The overall framework is made up of two components: one is the CNN for extracting primary features of the node images, while another is the graph neural network (GNN) for learning the features of edges and following updates. After this, first, the relationships among all the images are modeled by a fully-connected graph over learned CNN features. Then, a gated recurrent unit (GRU) takes the node and its corresponding information as inputs and outputs the updated node. As a result, the correlations among attributes can be modeled by using information from the neighbors of the node, as well as updating its state based on the previous state.

6.3 Adversarial Robustness in Facial Attribute Analysis

Adversarial images, which are generated from the network topology, training process, and hyperparameter variation, can be used as inputs of deep facial attribute analysis models with slight artificial perturbations. Then, by classifying the original inputs correctly and misclassifying the adversarial inputs, the robustness of models can be improved. Szegedy et al. [105] first propose that neural networks can be induced to misclassify an image by carefully chosen perturbations that are imperceptible to humans. After that, the study on adversarial images is entering the horizons of researchers.

Rozsa et al. [85] attempt to induce small artificial perturbations on existing misclassified inputs to correct the classification in facial attribute analysis. Specifically, the adversarial images are generated over a random subset of CelebA dataset via the fast flipping at-

tribute (FFA) technique. FFA algorithm leverages the back-propagation of a Euclidean loss without ground-truth labels to generate adversarial images by flipping the binary decision of the deep network. Separate facial attribute networks are trained for each attribute and tested the robustness of these networks by FFA. The experiments demonstrate that although FFA can create more adversarial images than the related fast gradient sign (FGS) algorithm [27], the few attributes can be affected when flipping the targeted attributes.

FFA algorithm flipping attributes enlightens the task of attribute anonymity, which conceals specific facial attributes that an individual does not want to share. During this process, the rest attributes should be maintained and the visual quality of images should be pledged. Chhabra et al. [13] achieve this basic target by virtue of the adversarial perturbation. An algorithm is derived to partition pre-defined attributes into a preservation set and a suppression set. Then, the images with adversarial perturbation can be generated based on these attribute sets. As a result, the prediction of a specific attribute from the true category can be classified to a different target category.

To sum up, the study of adversarial robustness contributes to improving the representational stability of current facial attribute estimation algorithms. Also, due to the attack of adversarial examples, deep facial attribute estimation models become more robust to achieve impressive performance.

7 Challenges and Opportunities

As we discussed before, facial attribute analysis is an important computer vision task both in theory and for real-world applications [99, 69, 43]. Firstly, we gave a description of what the facial attribute analysis is and what its two main sub-issues cover, i.e., facial attribute estimation (FAE) and facial attribute manipulation (FAM). Secondly, according to the general face image processing process, we reviewed the evolutions of both two sub-tasks and illustrated their pipelines respectively. Thirdly, we summarized the public databases as well as commonly used metrics. Fourthly, we introduced state-of-the-art algorithms of deep FAE and FAM with respective taxonomies, where part-based methods and holistic methods for FAE and model-based methods and extra condition-based methods for FAM. It is our firm belief that such taxonomies help us to understand the facial attribute analysis task better.

Despite the promising performance achieved by massive current algorithms in the field of deep facial attribute analysis, there are still several challenging issues that deserve more attention to be tackled. Therefore, in this section, we discuss challenges and future trends in both deep FAE and FAM respectively, from the perspectives of databases, algorithms, as well as the real-world applications.

7.1 Discussion of Facial Attribute Estimation

7.1.1 Data

The development of deep neural networks makes facial attribute estimation a data-driven task. That means a large number of samples are required for training deep neural networks sufficiently to capture attribute-relevant facial details. However, current facial attribute databases get trapped with the problem of insufficient data, which is a primary and major challenge.

Taking two datasets that are frequently used as examples, i.e., CelebA and LFWA, we provide the analysis of some problems in existing facial attribute databases.

Firstly, from the perspective of data sources, CelebA collects face data and attribute labels from the celebrities, while the samples of LFWA come from the news online. There is no doubt that these databases are inherently biased and do not match the general data distributions in the real world. For example, the *bald* attribute corresponds to a small number of samples in CelebA, but in the real world, it is a very common attribute among ordinary people. Hence, more complementary facial attribute datasets that cover more real-world scenarios and a wide range of facial attributes need to be constructed in the future. We anticipate that such effort in databases would significantly reduce the over-fitting for deep neural networks in the future.

Secondly, insufficient data lead to the data imbalance problem. We discuss such problem from two aspects: the imbalance in the distribution and the imbalance in the category. The imbalance in distribution is defined as the domain adaption problem, and the imbalance in the category is called class-imbalanced issue that has been discussed before. The analysis of the two imbalanced problems is provided below.

When predicting attributes from the datasets that collect samples from other data sources, e.g., the internet or videos, even any datasets that share different distributions with the training datasets, domain adaption becomes a new challenge. The discrepancies between data distributions would hurt the generalizability over unseen test data and lead to significant performance deterioration. Few studies consider the domain adaption issue in facial attribute estimation, bringing out an open issue for the future works.

Considering the class-imbalance issue, the imbalance on specific attributes can lead to a bias of clas-

sification performance, when features of minority attributes cannot be learned over sufficient samples. As we mentioned above, there have been a small amount of literature focusing on this issue in the field of facial attribute estimation, data pre-processing and cost-sensitive learning, which are typical algorithms for traditional class-imbalance learning, are still the keys to deal with the imbalanced problem for future deep facial attribute estimation.

7.1.2 Algorithms

Two types of deep FAE methods develop in parallel, i.e., part-based methods and holistic methods, while the former pays more attention to locate attributes, and the latter focuses more on modeling attribute relationships. We below provide the main challenges from the perspective of algorithms and analyze the future trends for both two types of methods, respectively.

For the part-based methods, earlier methods draw support from existing part detectors to discover facial components. However, these detected parts of faces are coarse and attribute-independent, as they only distinguish the whole face with the other face-irrelevant part, such as the background in an image. Considering that existing detectors are not customized for the facial attribute estimation, some researchers begin to seek help from other face-related auxiliary tasks, which focus more on facial details other than the whole face. There are also some studies utilizing labeled key points to partition facial regions. However, well-labeled facial images are not always available in the real-world applications, and the performance of auxiliary tasks would limit the accuracies of following classification tasks.

In the future, we believe that there is an urgent need for an end-to-end strategy, which learns attribute-relevant regions of the whole images, as well as predicts corresponding attributes over these regions in a unified framework. Although Ding et al. [16] have attempted to tackle this, learning a region for each attribute is unnecessary and computationally expensive, when several attributes might show up in the same region. How to model attribute relationships while locating their positions is a challenge for future part-based methods.

Besides, part-based methods show great superiority when dealing with the data under the in-the-wild environmental conditions, such as illumination variations, occlusions, and non-frontal faces. Through learning the locations of different attributes, part-based methods can integrate the information from non-occluded areas to predict attributes in occlusion areas. Mahbub et al. [72] deal with this issue by partitioning facial parts manually according to key points. However, such annotations are not always available, so trying to integrate these non-occluded areas adaptively is becoming a future trend. Besides, Mahbub et al. [72] test their model by adding occlusion manually, which is not normative in terms of the test protocol. Therefore, the lack of data under the in-the-wild conditions is still a challenge for training deep neural networks to solve facial attribute estimation in the wild environment.

As for holistic methods, state-of-the-arts design networks with different architectures for sharing common features and learning attribute-specific features at different layers. However, these methods define attribute relationships to design networks by grouping attributes manually, rather than learning such relationships adaptively. Needless to say, extra prior information given by researchers is introduced, and in the meantime, different individuals might partition different attribute groups according to locations or semantic. It is hard to determine that these current facial attribute groups are suitable and optimal. How to discover attribute relationships adaptively without given prior information artificially should be the focus of future works.

In addition, facial attributes have been taken as auxiliary and complementary information served for many face-related tasks, such as face recognition, face detection, and facial landmark localization. Despite of the promising performance achieved by pure facial attribute estimation over face images, joint and incorporative learning of these relevant tasks can further enhance their respective robustness and performance by discovering complementary information among them. For example, considering the inherent dependencies of face-related tasks, Zhuang et al. [126] design a cascaded CNN for simultaneously learning face detection, facial landmark localization, and facial attribute estimation under a multi-task framework to improve the performance of each task. They further attempt to joint face recognition with facial attribute estimation when taking the relationship between identities and attributes into account. Therefore, it is reasonable to think that the combination of different face-related tasks is becoming a promising research direction due to the complementarity among them.

7.1.3 Applications

Current researches mainly focus on the image data for predicting facial attributes, while the video data that contains a more extensive collection of images of the same person from varied viewpoints is significantly ignored. Earlier work [108] creates Ego-Humans dataset that draws out images from videos, where casual walkers are tracked in New York City throughout two months.

Nevertheless, Ego-Humans dataset extracts five million face pairs along with their same or not same labels with their weather and location contexts, which facilitates facial attribute prediction through additional features related to location and weather information. For example, sunny days encourage people to wear sunglasses and individuals in India Square are more likely to come from South Asia. Note that this work does not utilize the video data itself but extract it into image data, so that different viewpoints from the same person are not be used to assist facial attribute estimation.

On the one hand, such viewpoint diversification helps to learn richer features from the same person, as well as maintain identity-attribute consistency that aligns attributes of the same identity. However, on the other hand, attribute inconsistency becomes another new challenge where images of different viewpoints might differ in attributes even from the same identity. For example, the side face images might not disagree with the front face images for the prediction of *high cheekbones*, as the side face images do not emphasize this attribute, and even human cannot discriminate it precisely.

Liu et al. [68] work on the aforesaid attribute inconsistency problem over multiple face images. Probabilistic confidence criterion is proposed to address such inconsistency. Specifically, this criterion first extracts the most confident image for each subject, and then, it chooses the result corresponding to the highest confidence as the final prediction of each attribute with respect to each subject.

However, the study for this attribute inconsistency issue in the video data application still has a lacuna to be filled. Hence, we anticipate that more complementary algorithms and facial attribute video databases will be developed synergistically in the future.

Besides, nowadays, many devices in the real world contain enormous valuable personal information of users, such as bank accounts and personal emails. These personal details make these devices become the targets of various attacks. Hence, biological characteristics, such as fingerprints and irises, have been widely used as the passwords of these devices for further protecting the privacy information of users. It is the technique that we call as biometric identification. Moreover, more and more biometrics-based algorithms have emerged as a solution for continuous authentication on these devices. Many researchers have committed to designing active authentication algorithms based on face biometrics. For example, studies in [23, 28, 30] detect faces through camera sensor images and further extract low-level features for the authentication of smartphone users.

Inspired by these methods, we consider that facial attributes contain more detailed characteristics than the full face identification, and much more attention should be paid on attribute information for further advancing the progress of biometric identification on mobile devices. Samangouei et al. [88] have made an attempt on active authentication of mobile devices by virtue of facial attributes. A set of binary attribute classifiers are trained to estimate whether attributes are present in images of the current user in a mobile device. As a consequence, the authentication can be implemented by comparing the recognized attributes with the original enrolled ones.

However, Samangouei et al. [88] extract traditional features, such as LBP, which are not task-specific for facial attribute estimation and less discriminative than deep features. Besides, the binary attribute classifiers are also traditional SVMs. To some extent, the use of these traditional features and classifiers balances the verification accuracy and mobile performance, while some methods with good accuracies might have tremendous computation or memory costs. Hence, future challenges mainly lie in two aspects: the first is how to better apply facial attributes for the mobile device authentication, and the second is how to explore more discriminative deep features and classifiers under the constraints of the tradeoff between the verification accuracy and mobile performance.

7.2 Discussion of Facial Attribute Manipulation

7.2.1 Data

Generally, facial attribute manipulation is a conditional generative task, which synthesizes images according to given conditions. Note that the conditions that we discuss here are different from the conditions we mentioned when we create a taxonomy of state-of-the-art facial attribute manipulation methods. The conditions here denote broader concepts, which contain given input images, i.e., model-based methods, and other additional extra conditions, such as attribute vectors and reference exemplars. Compared with unconditional models that synthesize images only from random noises, facial attribute manipulation is a more controllable process. Moreover, facial attribute manipulation also can be determined as an image-to-image translation issue, which typically learns feature mappings from one image domain to another or discovers joint representations across different image domains, i.e., multi-domain transfer. That means many image-to-image translation algorithms can be directly used for facial attribute manipulation. However, in this paper, we only list studies that specifically do experiments on CelebA and LFWA facial attribute databases.

From the perspective of data, note that currently FAM methods only edit several conspicuous attributes, such as *smile*, *glasses*, *gender*, and *hair color*, several attributes that represent high-level semantic or abstractive details are not manipulated, such as *attractive*, *shadow* and *blurry*. To some extent, that is because it is difficult to define explicit concepts of these attributes since original images are labeled over these attributes subjectively. However, the synthesis of these abstractive attributes means a lot for the beauty makeup application. Moreover, *heavy makeup* and *age* attributes have become essential branches of image transformation, i.e., makeup manipulation [62, 9] and face aging [103], which are more tough tasks and attract the attention of many researchers. We believe they will have been being hot research topics in the future.

Besides, there is still a challenge to face: facial attribute manipulation in the video data still has not been studied, which is a more difficult issue as video images are dynamic. We believe this is because of the lack of relevant data and expect that more focus can be shifted to this field in the future.

Further, from the perspective of metrics, as we mentioned in Section 3, existing methods either evaluate generated images by statistical surveys or get help from other face-related tasks, such as attribute estimation and landmark detection. Unified and standard metric systems have not yet formed in terms of qualitative and quantitative analysis. We expect that the metrics of deep FAM methods could be well developed and form a unified rule in the future.

7.2.2 Algorithms

State-of-the-art deep FAM methods can be grouped into two categories: model-based methods and extra condition-based methods, while model-based methods tackle an attribute domain transfer issue and use the adversarial loss to supervise the process of image generation. Extra condition-based methods alter desired attributes with given conditional attributes concatenated with to-be-manipulated images in code spaces. The main difference between the two types of methods is whether extra conditions are required.

Model-based methods take no extra conditions as inputs, and one trained model only changes one corresponding attribute. This strategy is task-specific and helps to generate more photo-realistic images, but it is difficult to ensure attribute-irrelevant details unchanged due to its operation based on the whole image directly. Few methods have focused on this issue, except ResGAN proposed by Shen et al. [92]. However, ResGAN generates residual images for locating attribute-relevant regions under the sparsity constraint. Such constraint relies much on control parameters but not attributes themselves. Hence, how to design networks to synthesize desired photo-realistic attributes, as well as keep other attribute-irrelevant details unchanged, is a significant challenge in the future. In addition, as multi-domain transfer has become a hot research topic [64, 120], we expect that these novel domain transfer algorithms will migrate to the field of facial attributes for yielding more appealing performance.

Extra condition-based methods take attribute vectors or reference exemplars as conditions to edit facial attributes through changing values of attribute vectors, or latent codes of reference exemplars. One advantage of this type of methods is that multiple attributes can be manipulated simultaneously by altering multiple corresponding values of conditions. However, the concomitant disadvantage is also inevitable. That is these methods cannot change attributes continuously since values of attribute vectors are edited discretely. We believe that this shortcoming can be solved by interpolation [4] in the future. Besides, as we mentioned above, algorithms based on reference exemplars are becoming a promising research direction, since more specific details that appear in reference images can be explored to generate more realistic images compared with merely altering attribute vectors manually.

7.2.3 Applications

Facial makeup is a sub-issue in facial attribute manipulation, which can be significantly applied to various mobile devices, e.g., beauty cameras. This task pays more attention to facial details related to makeup, such as the types of eyeshadows and the colors of lipsticks. The focus of studies lies in facial makeup transfer and removal [9], where makeup transfer aims to map one makeup style to another for generating different makeup styles [62], and makeup removal performs an opposite process which cleans off the existed makeup and provides support to makeup-invariant face verification.

As an essential branch of facial attribute manipulation, facial makeup transfer and removal algorithms follow the general strategy of deep FAM by taking the makeup transfer as a domain transfer task, or a style transfer task among different makeup style images or makeup images and non-make up images. Even recently, facial makeup transfer based on reference exemplars becomes a popular topic of studies. For example, BeautyGAN [62] transfers makeup through incorporating domain transfer strategies with reference exemplars transfer. Specifically, discriminators distinguish generated images from the real ones at the domain level, while the

pixel-level histogram loss operates on separate local facial regions at the reference exemplar level. In this way, delicate makeup information can be learned and transformed to generate photo-realistic makeup images with different styles.

In addition, note that existing methods only work at the restricted range of resolutions. This constraint gives an opportunity to combine face super resolution with facial attribute manipulation. For example, Lu et al. [68] propose a conditional version of CycleGAN [125] to generate face images under the guidance of attributes for face super resolution. Taking a pair of low/high resolution faces as input, as well as an attribute vector that extracted from the high resolution one, attribute-guided conditional CycleGAN learns to generate a high-resolution version of the original low-resolution image, conditioned on attributes of the original high-resolution image. Besides, Dorta et al. [19] apply smooth warp fields to GANs for manipulating face images with very high resolutions through a deep network at a lower resolution. All these schemes inspire us to bring state-of-the-art face super resolution methods into facial attribute manipulation for achieving a win-win situation in the future.

So far, we have provided some possible challenges to be faced in deep FAE and FAM, and in the following, we will discuss the relationship between the two issues and how they assist each other.

For facial attribute estimation, facial attribute manipulation can be taken as a vital scheme of data augmentation, where generated facial attribute images can significantly increase the amount of data used to train deep neural networks. Further, overfitting could be reduced so that the accuracy of attribute prediction can be improved. When it comes to facial attribute manipulation, facial attribute estimation can be a significant quantitative performance evaluation criterion, where the accuracy gap between real images and generated images can be used to reflect the performance of facial attribute manipulation algorithms. However, despite of the mutual assistance, there are still some issues to be tackled for the two tasks.

Firstly, generated facial attribute images may not contain too much delicate facial information. That means there is still a gap between real and augmented generated images, and such a gap might damage the performance of estimation. Hence, how to close this gap can be an essential research direction for data augmentation in the future. Secondly, due to the indirect assessment of facial attribute manipulation, the performance of facial attribute estimation has side effects on the manipulation task. Therefore, how to balance the metric with the performance of prediction is another challenge. We expect that facial attribute estimation and facial attribute manipulation can strengthen their cooperation for significantly improving the performance of each other in the near future.

8 Conclusion

As one type of important semantic features describing visual properties of face images, facial attributes have received considerable attention in computer vision field. The analyses targeting for facial attributes, including facial attribute estimation (FAE) and facial attribute manipulation (FAM), have improved the performance of massive real-world applications. This paper provides a comprehensive review of recent advances in both deep learning based FAE and FAM, respectively. The commonly used databases and metrics are summarized, and the taxonomies of state-of-the-arts over both two issues have been created respectively, together with their pros and cons. Besides, future challenges and opportunities are highlighted in data, algorithms, and applications, respectively. We are looking forward to further studies that tackle these challenges to promote the development of deep face attribute analysis in the future.

Acknowledgements We thank the contributions that have made by pioneer researchers in the field of deep facial attribute analysis and other related fields. This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grant U1736119.

References

1. Amazon: Amazon mechanical turk. `https://www.mturk.com/`
2. Belghazi, M.I., Rajeswar, S., Mastropietro, O., Rostamzadeh, N., Mitrovic, J., Courville, A.: Hierarchical adversarially learned inference. arXiv preprint arXiv:1802.01071 (2018)
3. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 955–962. IEEE (2013)
4. Berthelot, D., Raffel, C., Roy, A., Goodfellow, I.: Understanding and improving interpolation in autoencoders via an adversarial regularizer. arXiv preprint arXiv:1807.07543 (2018)
5. Bourdev, L., Maji, S., Malik, J.: Describing people: A poselet-based approach to attribute classification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1543–1550. IEEE (2011)
6. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1365–1372. IEEE (2009)

7. Cao, J., Hu, Y., Yu, B., He, R., Sun, Z.: Load balanced gans for multi-view face image synthesis. arXiv preprint arXiv:1802.07447 (2018)
8. Cao, J., Li, Y., Zhang, Z.: Partially shared multi-task convolutional neural network with local constraint for face attribute learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4290–4299 (2018)
9. Chang, H., Lu, J., Yu, F., Finkelstein, A.: Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 40–48 (2018)
10. Chen, J.C., Ranjan, R., Sankaranarayanan, S., Kumar, A., Chen, C.H., Patel, V.M., Castillo, C.D., Chellappa, R.: Unconstrained still/video-based face verification with deep convolutional neural networks. International Journal of Computer Vision (IJCV) **126**(2-4), 272–291 (2018)
11. Chen, L., Zhang, Q., Li, B.: Predicting multiple attributes via relative multi-task learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1027–1034 (2014)
12. Chen, X., Duan, Y., Houthooft, R., Schulman, J., Sutskever, I., Abbeel, P.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), pp. 2172–2180 (2016)
13. Chhabra, S., Singh, R., Vatsa, M., Gupta, G.: Anonymizing k-facial attributes via adversarial perturbations. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), pp. 656–662 (2018)
14. Choi, Y., Choi, M., Kim, M.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8789–8797 (2018)
15. Cortes, C., Vapnik, V.: Support-vector networks. Machine learning **20**(3), 273–297 (1995)
16. Ding, H., Zhou, H., Zhou, S.K., Chellappa, R.: A deep cascade network for unaligned face attribute classification. arXiv preprint arXiv:1709.03851 (2017)
17. Doersch, C.: Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908 (2016)
18. Dong, Q., Gong, S., Zhu, X.: Class rectification hard mining for imbalanced deep learning. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1869–1878. IEEE (2017)
19. Dorta, G., Vicente, S., Campbell, N.D., Simpson, I.: The gan that warped: Semantic attribute editing with unpaired data. arXiv preprint arXiv:1811.12784 (2018)
20. Egger, B., Schönborn, S., Schneider, A., Kortylewski, A., Morel-Forster, A., Blumer, C., Vetter, T.: Occlusion-aware 3d morphable models and an illumination prior for face image analysis. International Journal of Computer Vision (IJCV) pp. 1–19 (2018)
21. Fan, Q., Gabbur, P., Pankanti, S.: Relative attributes for large-scale abandoned object detection. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2736–2743 (2013)
22. Fang, Y., Yuan, Q.: Attribute-enhanced metric learning for face retrieval. EURASIP Journal on Image and Video Processing **2018**(1), 44 (2018)
23. Fathy, M.E., Patel, V.M., Chellappa, R.: Face-based active authentication on mobile devices. In: Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1687–1691. IEEE (2015)
24. Gkioxari, G., Girshick, R., Malik, J.: Actions and attributes from wholes and parts. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2470–2478. IEEE (2015)
25. Gonzalez-Garcia, A., Modolo, D., Ferrari, V.: Do semantic parts emerge in convolutional neural networks? International Journal of Computer Vision (IJCV) **126**(5), 476–494 (2018)
26. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NIPS), pp. 2672–2680 (2014)
27. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: Proceedings of the International Conference on Learning Representations(ICLR) (2015)
28. Günther, M., Costa-Pazo, A., Ding, C., Boutellaa, E., Chiachia, G., Zhang, H., de Assis Angeloni, M., Štruc, V., Khoury, E., Vazquez-Fernandez, E., et al.: The 2013 face recognition evaluation in mobile environment. In: Proceedings of the International Conference on Biometrics (ICB), pp. 1–7. IEEE (2013)
29. Günther, M., Rozsa, A., Boult, T.E.: AFFACT: alignment-free facial attribute classification technique. In: Proceedings of the IEEE International Joint Conference on Biometrics (IJCB), pp. 90–99. IEEE (2017)
30. Hadid, A., Heikkila, J., Silvén, O., Pietikainen, M.: Face and eye detection for person authentication in mobile phones. In: Proceedings of the ACM/IEEE International Conference on Distributed Smart Cameras, pp. 101–108. IEEE (2007)
31. Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., Bing, G.: Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications **73**, 220–239 (2017)
32. Han, H., Jain, A.K., Shan, S., Chen, X.: Heterogeneous face attribute estimation: A deep multi-task learning approach. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2017)
33. Hand, E.M., Castillo, C.D., Chellappa, R.: Doing the best we can with what we have: Multi-label balancing with selective learning for attribute prediction. In: Proceedings of the Conference on Artificial Intelligence (AAAI), pp. 6878–6885 (2018)
34. Hand, E.M., Chellappa, R.: Attributes for improved attributes: a multi-task network utilizing implicit and explicit relationships for facial attribute classification. In: Proceedings of the Conference on Artificial Intelligence(AAAI), pp. 4068–4074 (2017)
35. He, D., Xia, Y., Qin, T., Wang, L., Yu, N., Liu, T., Ma, W.Y.: Dual learning for machine translation. In: Advances in Neural Information Processing Systems (NIPS), pp. 820–828 (2016)
36. He, K., Fu, Y., Xue, X.: A jointly learned deep architecture for facial attribute analysis and face detection in the wild. arXiv preprint arXiv:1707.08705 (2017)
37. He, K., Fu, Y., Zhang, W., Wang, C., Jiang, Y.G., Huang, F., Xue, X.: Harnessing synthesized abstraction images to improve facial attribute recognition. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2018)
38. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition(CVPR), pp. 770–778 (2016)
39. He, R., Tan, T., Davis, L., Sun, Z.: Learning structured ordinal measures for video based face recognition. Pattern Recognition **75**, 4–14 (2018)
40. He, R., Wu, X., Sun, Z., Tan, T.: Wasserstein cnn: Learning invariant features for nir-vis face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2018)
41. He, Z., Zuo, W., Kan, M., Shan, S., Chen, X.: Arbitrary facial attribute editing: Only change what you want. arXiv preprint arXiv:1711.10678 (2017)
42. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NIPS), pp. 6626–6637 (2017)
43. Hu, Y., Wu, X., Yu, B., He, R., Sun, Z.: Pose-guided photorealistic face rotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
44. Huang, C., Li, Y., Change Loy, C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5375–5384 (2016)
45. Huang, C., Li, Y., Loy, C.C., Tang, X.: Deep imbalanced learning for face recognition and attribute prediction. arXiv preprint arXiv:1806.00194 (2018)
46. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
47. Huang, H., Li, Z., He, R., Sun, Z., Tan, T.: Introvae: Introspective variational autoencoders for photographic image synthesis. arXiv preprint arXiv:1807.06358 (2018)
48. Huang, H., Song, L., He, R., Sun, Z., Tan, T.: Variational capsules for image analysis and synthesis. arXiv preprint arXiv:1807.04099 (2018)
49. Kalayeh, M.M., Gong, B., Shah, M.: Improving facial attribute prediction using semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4227–4235. IEEE (2017)
50. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1867–1874 (2014)
51. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
52. Kumar, N., Belhumeur, P., Nayar, S.: Facetracer: A search engine for large collections of images with faces. In: European Conference on Computer Vision (ECCV), pp. 340–353. Springer (2008)
53. Kumar, N., Berg, A., Belhumeur, P.N., Nayar, S.: Describable visual attributes for face verification and image search. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **33**(10), 1962–1977 (2011)
54. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 365–372. IEEE (2009)
55. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 951–958. IEEE (2009)
56. Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., et al.: Fader networks: Manipulating images by sliding attributes. In: Advances in Neural Information Processing Systems (NIPS), pp. 5967–5976 (2017)
57. Larsen, A.B.L., Sønderby, S.K., Larochelle, H., Winther, O.: Autoencoding beyond pixels using a learned similarity metric. In: Proceedings of the IEEE International Conference on Machine Learning (ICML), pp. 1558–1566 (2016)
58. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European Conference on Computer Vision (ECCV), pp. 679–692. Springer (2012)
59. Li, H.Y., Dong, W.M., Hu, B.G.: Facial image attributes transformation via conditional recycle generative adversarial networks. Journal of Computer Science and Technology **33**(3), 511–521 (2018)
60. Li, J., Zhao, F., Feng, J., Roy, S., Yan, S., Sim, T.: Landmark free face attribute prediction. IEEE Transactions on Image Processing **27**(9), 4651–4662 (2018)
61. Li, M., Zuo, W., Zhang, D.: Deep identity-aware transfer of facial attributes. arXiv preprint arXiv:1610.05586 (2016)
62. Li, T., Qian, R., Dong, C., Liu, S., Yan, Q., Zhu, W., Lin, L.: Beautygan: Instance-level facial makeup transfer with deep generative adversarial network. In: Proceedings of the ACM Multimedia Conference on Multimedia Conference (ACMMM), pp. 645–653. ACM (2018)
63. Li, Y., Wang, R., Liu, H., Jiang, H., Shan, S., Chen, X.: Two birds, one stone: Jointly learning binary code for large-scale face image retrieval and attributes prediction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3819–3827. IEEE (2015)
64. Liu, A.H., Liu, Y.C., Yeh, Y.Y., Wang, Y.C.F.: A unified feature disentangler for multi-domain image translation and manipulation. In: Advances in Neural Information Processing Systems (NIPS), pp. 2591–2600 (2018)
65. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in Neural Information Processing Systems (NIPS), pp. 700–708 (2017)
66. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 3730–3738 (2015)
67. Lu, Y., Kumar, A., Zhai, S., Cheng, Y., Javidi, T., Feris, R.: Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, p. 6 (2017)
68. Lu, Y., Tai, Y.W., Tang, C.K.: Attribute-guided face generation using conditional cyclegan. In: European Conference on Computer Vision (ECCV), pp. 293–308. Springer (2018)
69. Lu, Z., Hu, T., Song, L., Zhang, Z., He, R.: Conditional expression synthesis with face parsing transformation. In: Proceedings of the ACM International Conference on Multimedia (ACMMM), pp. 1083–1091. ACM (2018)
70. Luo, P., Wang, X., Tang, X.: A deep sum-product architecture for robust facial attributes analysis. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2864–2871. IEEE (2013)

71. Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., Van Gool, L.: Exemplar guided unsupervised image-to-image translation with semantic consistency. In: Proceedings of the International Conference on Learning Representations(ICLR) (2018)
72. Mahbub, U., Sarkar, S., Chellappa, R.: Segment-based methods for facial attribute detection from partial faces. arXiv preprint arXiv:1801.03546 (2018)
73. Meng, Z., Adluru, N., Kim, H.J., Fung, G., Singh, V.: Efficient relative attribute learning using graph neural networks. In: European Conference on Computer Vision (ECCV), pp. 552–567 (2018)
74. Miller, T.L., Berg, A.C., Edwards, J.A., Maire, M.R., White, R.M., Teh, Y.W., Learned-Miller, E., Forsyth, D.: Names and faces (2007)
75. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
76. Nguyen, H.M., Ly, N.Q., Phung, T.T.: Large-scale face image retrieval system at attribute level based on facial attribute ontology and deep neuron network. In: Asian Conference on Intelligent Information and Database Systems, pp. 539–549. Springer (2018)
77. Parikh, D., Grauman, K.: Relative attributes. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 503–510. IEEE (2011)
78. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: Proceedings of the British Machine Vision Conference 2015, (BMVC), vol. 1, p. 6 (2015)
79. Perarnau, G., van de Weijer, J., Raducanu, B., Álvarez, J.M.: Invertible conditional gans for image editing. arXiv preprint arXiv:1611.06355 (2016)
80. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8. IEEE (2007)
81. Ranjan, R., Patel, V.M., Chellappa, R.: 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 (TPAMI) (2017)
82. Ranjan, R., Sankaranarayanan, S., Castillo, C.D., Chellappa, R.: An all-in-one convolutional neural network for face analysis. In: Proceedings of the IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 17–24. IEEE (2017)
83. Rao, Y., Lu, J., Zhou, J.: Learning discriminative aggregation network for video-based face recognition and person re-identification. International Journal of Computer Vision (IJCV) pp. 1–18 (2018)
84. Rozsa, A., Günther, M., Rudd, E.M., Boult, T.E.: Are facial attributes adversarially robust? In: Processing of International Conference on Pattern Recognition (ICPR), pp. 3121–3127. IEEE (2016)
85. Rozsa, A., Günther, M., Rudd, E.M., Boult, T.E.: Facial attributes: Accuracy and adversarial robustness. Pattern Recognition Letters (2017)
86. Rudd, E.M., Günther, M., Boult, T.E.: Moon: A mixed objective optimization network for the recognition of facial attributes. In: European Conference on Computer Vision (ECCV), pp. 19–35. Springer (2016)
87. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 896–903 (2013)
88. Samangouei, P., Patel, V.M., Chellappa, R.: Facial attributes for active authentication on mobile devices. Image and Vision Computing **58**, 181–192 (2017)
89. Sandeep, R.N., Verma, Y., Jawahar, C.: Relative parts: Distinctive parts for learning relative attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3614–3621 (2014)
90. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 815–823 (2015)
91. Sethi, A., Singh, M., Singh, R., Vatsa, M.: Residual codean autoencoder for facial attribute analysis. Pattern Recognition Letters (2018)
92. Shen, W., Liu, R.: Learning residual images for face attribute manipulation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1225–1233. IEEE (2017)
93. Shi, H., Tao, L.: Fine-grained visual comparison based on relative attribute quadratic discriminant analysis. IEEE Transactions on Systems, Man, and Cybernetics: Systems (2018)
94. Shi, Z., Hospedales, T.M., Xiang, T.: Transferring a semantic representation for person re-identification and search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4184–4193. IEEE (2015)
95. Singh, K.K., Lee, Y.J.: End-to-end localization and ranking for relative attributes. In: European Conference on Computer Vision (ECCV), pp. 753–769. Springer (2016)
96. Smith, B.M., Zhang, L., Brandt, J., Lin, Z., Yang, J.: Exemplar-based face parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3484–3491. IEEE (2013)
97. Song, F., Tan, X., Chen, S.: Exploiting relationship between attributes for improved face verification. Computer Vision and Image Understanding **122**, 143–154 (2014)
98. Song, L., Cao, J., Song, L., Hu, Y., He, R.: Geometry-aware face completion and editing. In: Proceedings of the Conference on Artificial Intelligence(AAAI) (2019)
99. Song, L., Lu, Z., He, R., Sun, Z., Tan, T.: Geometry guided adversarial facial expression synthesis. In: Proceedings of the ACM International Conference on Multimedia (ACMMM), pp. 627–635. ACM (2018)
100. Song, L., Zhang, M., Wu, X., He, R.: Adversarial discriminative heterogeneous face recognition. In: Proceedings of the Conference on Artificial Intelligence(AAAI) (2018)
101. Sun, R., Huang, C., Shi, J., Ma, L.: Mask-aware photorealistic face attribute manipulation. arXiv preprint arXiv:1804.08882 (2018)
102. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems (NIPS), pp. 1988–1996 (2014)
103. Suo, J., Zhu, S.C., Shan, S., Chen, X.: A compositional and dynamic model for face aging. IEEE Transactions on Pattern Analysis and Machine Intelligence(TPAMI) **32**(3), 385–401 (2010)
104. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the Conference on Artificial Intelligence (AAAI), vol. 4, p. 12 (2017)

105. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. In: Proceedings of the International Conference on Learning Representations(ICLR) (2014)
106. Toderici, G., Omalley, S.M., Passalis, G., Theoharis, T., Kakadiaris, I.A.: Ethnicity-and gender-based subject retrieval using 3-d face-recognition techniques. International Journal of Computer Vision (IJCV) **89**(2-3), 382–391 (2010)
107. Tropp, J.A., Gilbert, A.C., Strauss, M.J.: Algorithms for simultaneous sparse approximation. part i: Greedy pursuit. Signal Processing **86**(3), 572–588 (2006)
108. Wang, J., Cheng, Y., Schmidt Feris, R.: Walk and learn: Facial attribute representation learning from egocentric video and contextual data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2295–2304 (2016)
109. Wang, Y., Wang, S., Qi, G., Tang, J., Li, B.: Weakly supervised facial attribute manipulation via deep adversarial network. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 112–121. IEEE (2018)
110. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
111. Wu, Y., Ji, Q.: Facial landmark detection: A literature survey. International Journal of Computer Vision (IJCV) pp. 1–28 (2017)
112. Xiao, F., Jae Lee, Y.: Discovering the spatial extent of relative attributes. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 1458–1466 (2015)
113. Xiao, T., Hong, J., Ma, J.: Elegant: Exchanging latent encodings with gan for transferring multiple face attributes. arXiv preprint arXiv:1803.10562 (2018)
114. Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: European Conference on Computer Vision (ECCV), pp. 776–791. Springer (2016)
115. Yue, Y., Finley, T., Radlinski, F., Joachims, T.: A support vector method for optimizing average precision. In: Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 271–278. ACM (2007)
116. Zhang, G., Kan, M., Shan, S., Chen, X.: Generative adversarial network with spatial attention for face attribute editing. In: European Conference on Computer Vision (ECCV), pp. 417–432 (2018)
117. Zhang, J., Shu, Y., Xu, S., Cao, G., Zhong, F., Qin, X.: Sparsely grouped multi-task generative adversarial networks for facial attribute manipulation. arXiv preprint arXiv:1805.07509 (2018)
118. Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1637–1644. IEEE (2014)
119. Zhang, S., He, R., Sun, Z., Tan, T.: Demeshnet: Blind face inpainting for deep meshface verification. IEEE Transactions on Information Forensics and Security (TIFS) **13**(3), 637–647 (2018)
120. Zhang, Y.: Xogan: One-to-many unsupervised image-to-image translation. arXiv preprint arXiv:1805.07277 (2018)
121. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 4352–4360 (2017)
122. Zhong, Y., Sullivan, J., Li, H.: Face attribute prediction using off-the-shelf cnn features. In: Proceedings of the IEEE International Conference on Biometrics (ICB), pp. 1–7. IEEE (2016)
123. Zhong, Y., Sullivan, J., Li, H.: Leveraging mid-level deep representations for predicting face attributes in the wild. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 3239–3243. IEEE (2016)
124. Zhou, S., Xiao, T., Yang, Y., Feng, D., He, Q., He, W.: Genegan: Learning object transfiguration and attribute subspace from unpaired data. arXiv preprint arXiv:1705.04932 (2017)
125. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2242–2251 (2017)
126. Zhuang, N., Yan, Y., Chen, S., Wang, H.: Multi-task learning of cascaded cnn for facial attribute classification. arXiv preprint arXiv:1805.01290 (2018)
127. Zhuang, N., Yan, Y., Chen, S., Wang, H., Shen, C.: Multi-label learning based deep transfer neural network for facial attribute classification. Pattern Recognition **80**, 225–240 (2018)