REVIEW

A survey on face data augmentation for the training of deep neural networks

Xiang Wang[1] · Kai Wang[1] · Shiguo Lian[2]

Abstract

The quality and size of training set have a great impact on the results of deep learning-based face-related tasks. However, collecting and labeling adequate samples with high-quality and balanced distributions still remains a laborious and expensive work, and various data augmentation techniques have thus been widely used to enrich the training dataset. In this paper, we review the existing works of face data augmentation from the perspectives of the transformation types and methods, with the state-of-the-art approaches involved. Among all these approaches, we put the emphasis on the deep learning-based works, especially the generative adversarial networks which have been recognized as more powerful and effective tools in recent years. We present their principles, discuss the results and show their applications as well as limitations. Different evaluation metrics for evaluating these approaches are also introduced. We point out the challenges and opportunities in the field of face data augmentation and provide brief yet insightful discussions.

Keywords Data augmentation · Face image transformation · Generative models

1 Introduction

In recent years, face studies in computer vision have shifted from seeking engineering features by hand to using deep learning approaches. As a result, data play a more important role, since the performance of deep neural network heavily depends on the amount and quality of the training data. The remarkable work by Facebook [129] and Google [116] demonstrated the effectiveness of large-scale datasets on obtaining high-quality trained model and revealed that deep learning strongly relies on large and complex training sets to generalize well in unconstrained settings. However, collecting and labeling a large quantity of real samples is widely recognized as laborious,

expensive and error prone, and existing datasets are usually lack of balance in data distribution.

To compensate the insufficient training data, domain adaptation transfers the knowledge learned from a source data distribution to a target data distribution [23–26]. However, it needs samples from target distributions before training and assumes the target distribution is fixed, which is restrictive in practical scenarios [138]. On the other hand, data augmentation provides an effective alternative, which we call "face data augmentation." It is a technology to enlarge the data size of training or testing by transforming collected real face samples or simulated virtual face samples. Figure 1 shows a schematic diagram of face data augmentation, which is our focus in this paper.

Assuming the original dataset is $\mathcal{S}$, face data augmentation can be represented by the following mapping:

$$\phi : \mathcal{S} \mapsto \mathcal{T} \tag{1}$$

where $\mathcal{T}$ is the augmented set of $\mathcal{S}$. Then the dataset is enlarged as the union of the original set and the augmented set:

$$\mathcal{S}' = \mathcal{S} \cup \mathcal{T} \tag{2}$$

✉ Xiang Wang
wangxiang8602@126.com

Kai Wang
kai.wang@cloudminds.com

Shiguo Lian
liansg@chinaunicom.cn

[1] CloudMinds Technologies Inc., Beijing, China
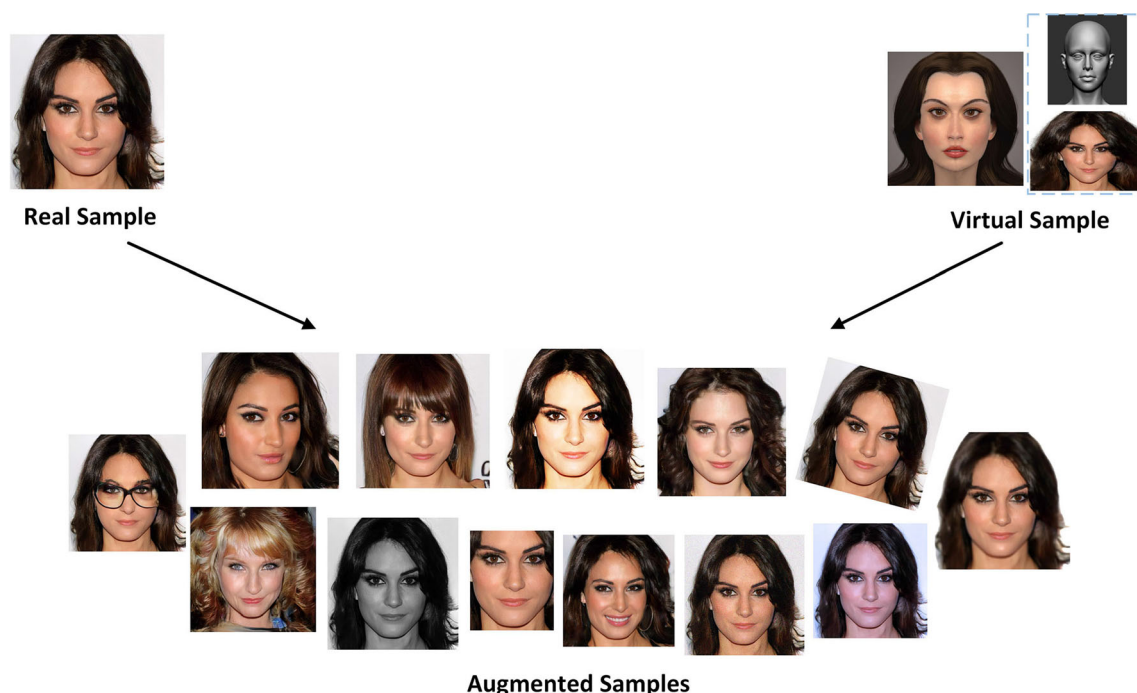
[2] China Unicom, Beijing, China

Fig. 1 Schematic diagram of face data augmentation. The real and augmented samples were generated by TL-GAN (transparent latent-space generative adversarial network) [40]

The direct motivation for face data augmentation is to overcome the limitation of existing data. Insufficient data amount or unbalanced data distribution will cause overfitting and over-parameterization problems, leading to an obvious decrease in the effectiveness of learning result.

Face data augmentation is fundamentally important for improving the performance of neural networks in the following aspects. (1) It is inexpensive to generate a huge number of synthetic data with annotations in comparison to collecting and labeling real data. (2) Synthetic data can be accurate, so it has ground-truth by nature. (3) If controllable generation method is adopted, faces with specific features and attributes can be obtained. (4) Face data augmentation has some special advantages, such as generating faces without self-occlusion [33] and balanced dataset with more intra-class variations [98].

At the same time, face data augmentation has limitations. (1) Some augmentation is challenging and no perfect solution exists so far, such as predicting face appearance in the presence of large pose variation, augmenting naturalistic faces with expressions, age progression and regression. (2) The domain gap between real and virtual should be considered if the synthetic data are not realistic enough or the synthetic data domain has different distribution to real data domain. That is why some researchers use domain adaption and transfer learning techniques to improve the utility of synthetic data [48, 83].

This paper aims to give an epitome of face data augmentation, especially on what can face data augmentation do and how to augment the existing face data, including both the traditional methods and the up-to-date approaches. In addition, we thoroughly discuss the challenges and open problems of face data augmentation for further research. Data augmentation has overlapped with data synthesis/generation, but differs with them in the point that the augmented data is generated based on existing data. In fact, many data synthesis techniques can be applied to data augmentation. Although some works were not designed for data augmentation, we also include them in this survey.

The remainder of the paper is organized as follows. The review of related works is given in Sect. 2. Then the transformation types of face data augmentation are elaborated in Sect. 3, and the commonly used methods are introduced in Sect. 4. Section 5 provides a description of the evaluation metrics. Section 6 presents some challenges and potential research directions. Discussions and conclusion are given in the last two sections, respectively.

2 Related work

This section reviews the existing works that have in-depth analysis and evaluation on data augmentation techniques. Masi et al. [98] discussed the necessity of collecting huge numbers of face images for effective face recognition and

proposed a synthesizing method to enrich the existing dataset by introducing face appearance variations for pose, shape and expression. Lv et al. [94] presented five data augmentation methods for face images, including landmark perturbation, hairstyle synthesis, glasses synthesis, poses synthesis and illumination synthesis. They tested these methods on different datasets for face recognition, and compared their performance. Taylor et al. [130] demonstrated the effectiveness of using basic geometric and photometric data augmentation schemes like cropping, rotation, etc., to help researchers find the most appropriate choice for their dataset. Wang et al. [139] compared traditional transformation methods with GANs (generative adversarial networks) to the problem of data augmentation in image classification. In addition, they proposed a network-based augmentation method to learn augmentations that best improve the classifier in the context of generic images, not face images. Kortylewski et al. [71] explored the ability of data augmentation to train deep face recognition systems with an off-the-shelf face recognition software and fully synthetic face images varying in pose, illumination and background. Their experiment demonstrated that synthetic data with strong variations performed well across different datasets even without dataset adaptation, and the domain gap between the real and the synthetic could be closed when using synthetic data for pre-training followed by fine-tuning. Li et al. [79] reviewed the research progress in the field of virtual sample generation for face recognition. They categorized the existing methods into three groups: construction of virtual face images based on face structure, perturbation and distribution function and sample viewpoint. Compared to the existing works, our survey covers a wider range of face data augmentation methods and contains the up-to-date researches. We introduce these researches in intuitive presentation level and deep method level.

3 Transformation types

In this section, we elaborate the transformation types, including the generic and face-specific transformations, for producing the augmented samples $\mathcal{T}$ in Eq. 1. The applications of some methods go beyond face data augmentation to other learning-based computer vision tasks. Usually, the generic methods transform the entire image and ignore high-level contents, while face-specific methods focus on facial attributes and are capable of transforming age, makeup, hairstyle, etc. Table 1 shows an overview of the commonly used face data transformations. All the illustrations for the facial attribute transformation are created by TL-GAN [40] except the age progression, which is created by FaceApp [32].

3.1 Geometric and photometric transformation

The generic data augmentation techniques can be divided into two categories: geometric transformation and photometric transformation. These methods have been adapted to various learning-based computer vision tasks.

Geometric transformation alters the geometry of an image by transferring image pixel values to new positions. This kind of transformation includes translation, rotation, reflection, flipping, zooming, scaling, cropping, padding, perspective transformation, elastic distortion, lens distortion, mirroring, etc. Some examples are illustrated in Fig. 2, which are created using imgaug [60]—a python library for image augmentation.
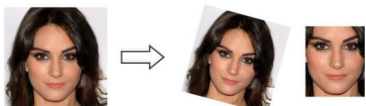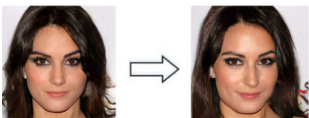
Photometric transformation alters the RGB channels by shifting pixel colors to new values, and the main approaches include color jittering, grayscaling, filtering, lighting perturbation, noise adding, vignetting, contrast adjustment, random erasing, etc. The color jittering method includes many different manipulations, such as inverting, adding, decreasing and multiplying. The filtering method includes edge enhancement, blurring, sharpening, embossing, etc. Some examples of the photometric transformations are shown in Fig. 3.

Wu et al. [143] adopted a series of geometric and photometric transformations to enrich the training dataset and prevent overfitting. Taylor et al. [130] evaluated six geometric and photometric data augmentation methods with baseline for the task of image classification. The six data augmentation methods included flipping, rotating, cropping, color jittering, edge enhancement and fancy PCA (adding multiples of principle components to the images). Their results indicated that data augmentation helped improving the classification performance of convolutional neural network (CNN) in all cases, and the geometric augmentation schemes outperformed the photometric schemes, whereas the theoretical proof remains the topic of ongoing research. Mash et al. [96] also benchmarked a variety of augmentation methods, including cropping, rotating, rescaling, polygon occlusion and their combinations, in the context of CNN-based fine-grain aircraft classification. The experimental results showed that flipping and cropping have more obvious improvement on the classifier performance than random scaling and occlusions, which is consistent with the demonstrations in other generic data augmentation works.

3.2 Hairstyle transfer

Although hair is not an internal component of human face, it affects face detection and recognition due to the occlusion and appearance variation of face it caused. The data

Table 1 An overview of transformation types

Transformation type		Illustration
Generic transformation	Geometric	
	Photometric	
Facial attribute transformation	Hairstyle	
	Makeup	
	Accessory	
	Pose	
	Expression	
	Age	

augmentation technique is used to generate face images with different hairs in color, shape and bang.

Kim et al. [66] transformed hair color using DiscoGAN, which was introduced to discover cross-domain relations with unpaired data. The same function was achieved by StarGAN [19] and AttGAN [46], whereas it could perform multi-domain translations using a single model. Besides the color, [67] proposed an unsupervised visual attribute transfer using reconfigurable generative adversarial network to change the bang. Lv et al. [94] synthesized images with various hairstyles based on hairstyle templates. Kemelmacher-Shlizerman et al. [63] presented a face synthesis system utilizing an Internet-based compositing method. With one or more photos of a person's face and a text query curly hair as input, the system could generate a series of new photos with the input persons identity and queried appearance.

Figure 4 shows some examples of hairstyle transfer created by AttGAN [46].

3.3 Facial makeup transfer

While makeup is a ubiquitous way to improve one's facial appearance, it increases the difficulty on accurate face recognition. Therefore, numerous samples with different makeup styles should be provided in the training data to make the algorithm be more robust. Facial makeup transfer aims to shift the makeup style from a given reference to another face while preserving the face identity. Common makeups include foundation, eye liner, eye shadow, lipstick, etc., and their transformations include applying,

Fig. 2 Geometric transformation examples created by imgaug [60]. The source image is from CelebA dataset [91]. From left to right and from top to bottom, the transformations are crop&pad, elastic distortion, scale, piecewise affine, translate, horizontal flip, vertical flip, rotate, perspective transformation and shear
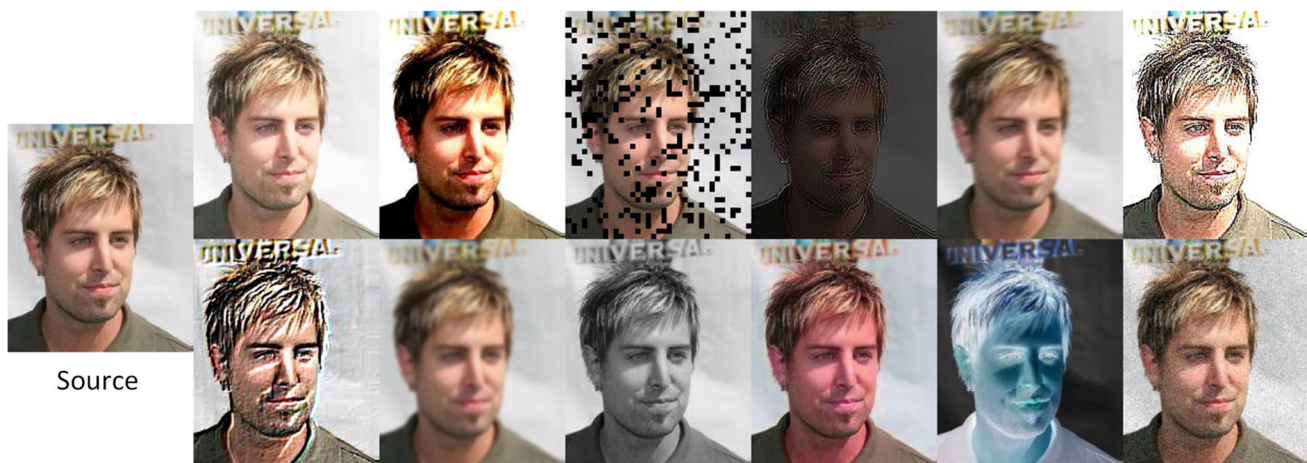

Fig. 3 Photometric transformation examples created by imgaug [60]. The source image is from CelebA dataset [91]. From left to right and from top to bottom, the transformations are brightness change, contrast change, coarse dropout, edge detection, motion blur, sharpen, emboss, Gaussian blur, hue and saturation change, invert and adding noise

removal and exchange. Most existing studies on automatic makeup transfer can be classified into two categories: traditional image processing approaches like gradient editing and alpha blending [42, 76, 104] and deep learning-based methods [2, 16, 82, 89].

Guo et al. [42] transferred face makeup from one image to another by decomposing face images into different layers, which were face structure, skin details and color. Then, each layer of the example and original images were combined to obtain a natural makeup effect through gradient editing, weighted addition and alpha blending. Similarly, [104] decomposed the reference and target images into large-scale layer, detail layer and color layer, where the makeup highlight and color information was transferred by Poisson editing, weighted means and alpha blending. Lee and Kang [76] presented a method of makeup applying

for different regions based on multiple makeup examples, e.g., the makeup of eyes and lip were taken from two references, respectively.

Traditional methods usually consider the makeup style as a combination of different components, and the overall output image usually looks unnatural with artifacts at adjacent regions [82]. In contrast, the end-to-end deep networks acting on entire image showed great advantage in terms of output quality and diversity. For example, Liu et al. [89] proposed a deep localized makeup transfer network to automatically generate faces with makeup by applying different cosmetics to corresponding facial regions in different manners. Alashkar et al. [2] presented an automatic makeup recommendation and synthesis system based on a deep neural network trained from pairwise of before–after makeup images united with artist
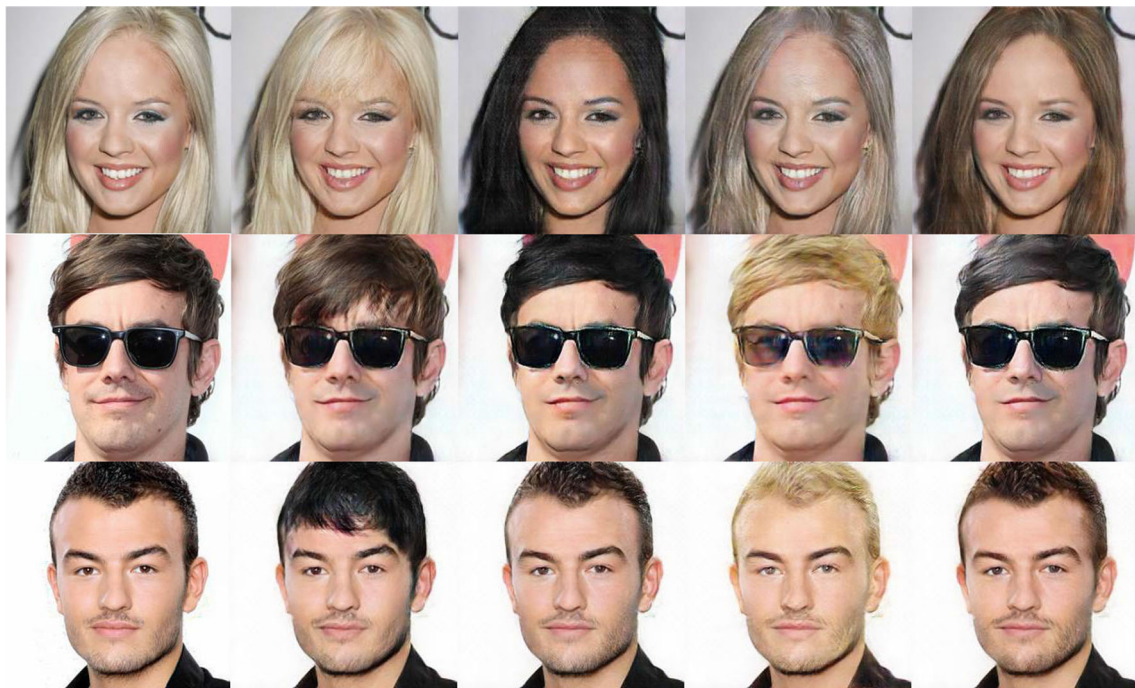
Fig. 4 Hairstyle transfer examples created by AttGAN [46]. From left to right, there are input images (from CelebA dataset [91]), transferred results for bangs, black hair, blond hair and brown hair

knowledge rules. Nguyen et al. [103] also proposed an automatic and personalized facial makeup recommendation and synthesis system. However, their system was realized based on a latent SVM model describing the relations among facial features, facial attributes and makeup attributes. With the rapid development of generative adversarial networks, they have been widely used in makeup transfer. Chang et al. [16] used cycle-consistent generative adversarial networks to simultaneously learn a makeup transfer function and a makeup removal function, such that the output of the network after a cycle should be consistent with the original input and no paired training data was needed. Similarly, BeautyGAN [82] applied a dual generative adversarial network. However, they further improved it by incorporating global domain-level loss and local instance-level loss to ensure an appropriate style transfer for both the global (identity and background) and independent local regions (like eyes and lip).

Figure 5 presents some makeup transfer examples created by StarGAN [19]. We need a statement that these resulting images do not represent the quality of the network model because the quality of the training data we use has great impact on the transformation results.

3.4 Accessory removal and wearing

The removal and wearing of accessories include those of the glasses, earrings, nose ring, nose stud, lip ring, etc.

Among all the accessories, glasses are mostly common seen, as they are worn for various purposes including vision correction, bright sunlight prevention, eye protection, beauty, etc. Glasses could significantly affect the accuracy of face recognition as they usually cover a large area of human faces.

Lv et al. [94] synthesized glasses-wearing images using template-based method. Guo et al. [43] fused virtual eyeglasses with face images by augmented reality technique. The InfoGAN proposed in [18] learned disentangled representations of faces in a completely unsupervised manner and was able to modify the presence of glasses. Shen et al. [117] proposed a face attribute manipulation method based on residual image, which was defined as the difference between the input image and the desired output image. They adopted two inverse image transformation networks to produce residual images for glasses wearing and removing.

Figure 6 presents some examples of glasses wearing and removal.

3.5 Pose transformation

The large pose discrepancy of head in the wild proposes a big challenge in face detection and recognition tasks, as self-occlusion and texture variation usually occur when head pose changes. Therefore, a variety of pose-invariant methods were proposed, including data augmentation for

Fig. 5 Facial makeup transfer examples created by StarGAN [19]. The images on the top row are the input (from CelebA dataset [91]), and the second row is the transformation result

Fig. 6 Glasses wearing and removal examples created by AttGAN [46]. The top row is the input images (from CelebA dataset [91]), the second row is the transformation result
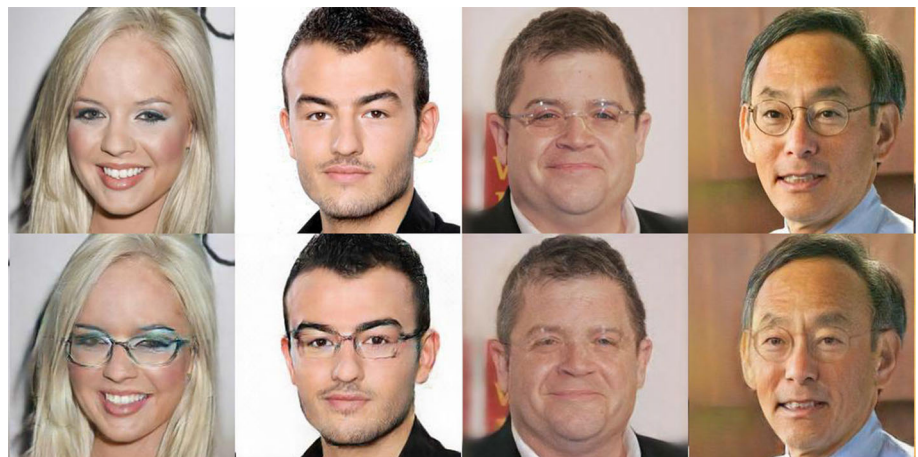
different poses. Since existing datasets mainly consist of near-frontal faces, which contradicts the condition for unconstrained face recognition.

Feng et al. [34] proposed a 2D profile face generator, which produced out-of-plane pose variations based on a PCA-based 2D shape model. Meanwhile, many works used 3D face models for face pose translation [21, 44, 73, 94, 98, 165]. Facial texture is an important component of 3D face model and directly affects the reality of generated faces. Deng et al. [27] proposed a generative adversarial network (UV-GAN) to complete the facial UV map and recover the self-occluded regions. In their experiment, virtual instances under arbitrary poses were generated by attaching the completed UV map to the fitted 3D face mesh. In another way, Zhao et al. [159, 160] used network to enhance the realism of synthetic face images generated by 3D morphable model.

Generative models are pervasively employed by recent works to synthesize faces with arbitrary poses. Van den Oord et al. [136] applied a conditional PixelCNN architecture to generate new portraits with different poses conditioned on pose embeddings. Wiles et al. [142]

introduced X2Face that could control a source face by a driving frame to produce a generated face with the identity of the source but the pose and expression of the other. Hu et al. [51] proposed Couple–Agent Pose-Guided Generative Adversarial Network (CAPG-GAN) to realize flexible face rotation of arbitrary head poses from a single image in 2D space. They employed facial landmark heatmap to encode the head pose to control the generator. Moniz et al. [102] presented DepthNets to infer the depth of facial keypoints from the input image and predict 3D affine transformations that maps the input face to a desired pose and facial geometry. Cao et al. [14] introduced Load Balanced Generative Adversarial Networks (LB-GAN) and decomposed the face rotation problem into two subtasks, which frontalize the face images first and rotate the front-facing faces later. Through their comparison experiment, LB-GAN performed better in identity preserving.

A special case of pose transformation is face frontalization. It is commonly used to increase the accuracy rate of face recognition by rotating faces to the front view, which are more friendly to the recognition model. Hassner et al. [45] rotated a single, unmodified 3D reference

face model for face frontalization. Taigman et al. [129] warped facial crops to frontal mode for accurate face alignment. TP-GAN [54] and FF-GAN [151] are two classic face frontalization methods based on GANs. TP-GAN is a Two-Pathway Generative Adversarial Network for photorealistic frontal view synthesis by simultaneously perceiving global structures and local details. FF-GAN incorporates 3DMM into the GAN structure to provide shape and appearance priors for fast convergence with less training data. Furthermore, Zhao et al. [158] proposed a Pose-Invariant Model (PIM) by jointly learning face frontalization and facial features end-to-end for pose-invariant face recognition.

3.6 Expression synthesis and transfer

The facial expression synthesis and transfer technique is used to enrich the expressions (happy, sad, angry, fear, surprise and disgust) of a given face and helps to improve the performance of tasks like emotion classification, expression recognition and expression-invariant face recognition. The expression synthesis and transfer methods can be classified into 2D geometry-based approach, 3D geometry-based approach and learning-based approach.

The 2D- and 3D-based algorithms emerged earlier than learning-based methods, whose evident advantage is that they do not need a large amount of training samples. The 2D-based algorithms transfer expressions relying on the geometry and texture features of the expressions in 2D space, such as the work from [147], while the 3D-based algorithms generate face images with various emotions from 3D face models or 3D face data. Agianpuye and Minoi [1] gave a comprehensive survey on the field of 3D facial expression synthesis. Until nowadays, "Blendshapes" model which was introduced in computer graphics remains the most prevalent approach for 3D expression synthesis and transfer [44, 65, 98, 131].

In recent years, a lot of learning-based methods were proposed for expression synthesis and transfer. Besides the use of CNN [35, 80, 136], wide application of generative models of autoencoders and GANs has begun. For example, Yeh et al. [150] combined the flow-based face manipulation with variational autoencoders to encode the flow from one expression to another over a low-dimensional latent space. Zhou et al. [162] proposed the conditional difference adversarial autoencoder (CDAAE), which can generate specific expression for unseen person with a target emotion or facial action unit (AU) label. On the other side, the GAN-based methods can be further divided into four categories according to the generative condition of expression generation: reference images [11, 167], emotion or expression codes [29, 152], action unit labels [109, 110] and geometry [111, 127, 144].

Zhu et al. [167] and Bao et al. [11] generated different emotions with reference and target input. Zhang et al. [152] generated different expressions under arbitrary poses conditioned by the expression and pose one-hot codes. Ding et al. [29] proposed an Expression Generative Adversarial Network (ExprGAN) for photorealistic facial expression editing with controllable expression intensity. They designed an expression controller module to generate expression code, which was a real-valued vector containing the expression intensity description. Pham et al. [109] proposed a weakly supervised adversarial learning framework for automatic facial expression synthesis based on continuous action unit coefficients. Pumarola et al. [110] also controlled the generated expression by AU labels and allowed a continuous expression transformation. In addition, they introduced an attention-based generator to promote the robustness of their model for distracting backgrounds and illuminations. Song et al. [127] proposed a Geometry-Guided Generative Adversarial Network (G2-GAN) to synthesize photorealistic and identity-preserving facial images in different expressions from a single image. They employed facial geometry (fiducial points) as the controllable condition to guide facial texture synthesis. Qiao et al. [111] used geometry (facial landmarks) to control the expression synthesis with a facial geometry embedding network and proposed a Geometry-Contrastive Generative Adversarial Network (GC-GAN) to transfer continuous emotions across different subjects even there were big shape difference. Furthermore, Wu et al. [144] proposed a boundary latent space and boundary transformer. They mapped the source face into the boundary latent space and transformed the source face's boundary to the target's boundary, which was the medium to capture facial geometric variances during expression transfer.

Figure 7 shows some examples created by AttGAN [46]. It can transfer neutral face into smile one and vice versa.

3.7 Age progression and regression

Age progression or face aging predicts one's future looks based on his current face, while age regression or rejuvenation estimates one's previous looks. All of them aim to synthesize faces of various ages and preserve personalized features at the same time. The generated face images enrich the data of individual subjects over a long range of age span, which enhances the robustness of the learned model to age variation.

The traditional methods of age transfer include the prototype-based method and model-based method. The prototype-based method creates average faces for different age groups, learns the shape and texture transformation between these groups and applies them to images for age

Fig. 7 Expression transformation examples created by AttGAN [46]. The top row is the input images (from CelebA dataset [91]), the second row is the transformation result

transfer [64]. However, personalized features on individual faces are usually lost in this method. The model-based method constructs parametric models of biological facial change with age, e.g., muscle, wrinkle, skin, etc. But such models typically suffer from high complexity and computational cost [156]. In order to avoid the drawbacks of the traditional methods, Suo et al. [128] presented a compositional dynamic model and applied a three level And-Or graph to represent the decomposition and the diversity of faces by a series of face component dictionaries. Similarly, Shu et al. [121] learned a set of age-group specific dictionaries and used a linear combination to express the aging process. In order to preserve the personalized facial characteristics, every face was decomposed into an aging layer and a personalized layer for consideration of both the general aging characteristics and the personalized facial characteristics.

More recent works applied GANs with encoders for age transfer. The input images are encoded into latent vectors, transformed in the latent space and reconstructed back into images with a different age. Palsson et al. [105] proposed three aging transfer models based on CycleGAN. Wang et al. [140] proposed a recurrent face aging (RFA) framework based on recurrent neural network. Zhang et al. [156] proposed a conditional adversarial autoencoder (CAAE) for face age progression and regression, based on the assumption that the face images lay on a high-dimensional manifold, and the age transformation could be achieved by a traversing along a certain direction. Antipov et al. [4] proposed Age Conditional Generative Adversarial Network (Age-cGAN) for automatic face aging and rejuvenation, which emphasized on identity preserving and introduced an approach for the optimization of latent vectors. Wang et al. [141] proposed an Identity-Preserved Conditional Generative Adversarial Networks (IPCGANs). They

combined a Conditional Generative Adversarial Network with an identity-preserved module and an age classifier to generate photorealistic faces in a target age. Zhao et al. [157] proposed a GAN-like Face Synthesis sub-Net (FSN) to learn a synthesis function that can achieve both face rejuvenation and aging with remarkable photorealistic and identity-preserving properties without the requirement of paired data and the true age of testing samples. Zhu et al. [163] paid more attention to the aging accuracy and utilized an age estimation technique to control the generated face. Li et al. [81] introduced a Wavelet-domain Global and Local Consistent Age Generative Adversarial Network (WaveletGLCA-GAN) for age progression and regression. Liu et al. [90] pointed out that only identity preservation was not enough, especially for those models trained on unpaired face aging data. They proposed an attribute-aware face aging model with wavelet-based GANs to ensure attribute consistency.

3.8 Other styles transfer

In addition to the transformations summarized above, there are also some other types of transformations to enrich the face dataset, such as face illumination transfer [11, 21, 44, 73, 78, 94, 136, 168], gender transfer [19, 66, 80, 117], skin color transfer [19, 46], eye color transfer [18], eyebrows transfer [46, 74], mustache or beard transfer [46, 63], facial landmark perturbation [93], context and background change [9]. Figure 8 presents some examples of these transfer result.

In recent years, there have been more and more works trying to achieve multiple types of transformations through a unified neural network. For example, StarGAN [19] is able to perform expression, gender, age and skin color transformations with a unified model and a one-time
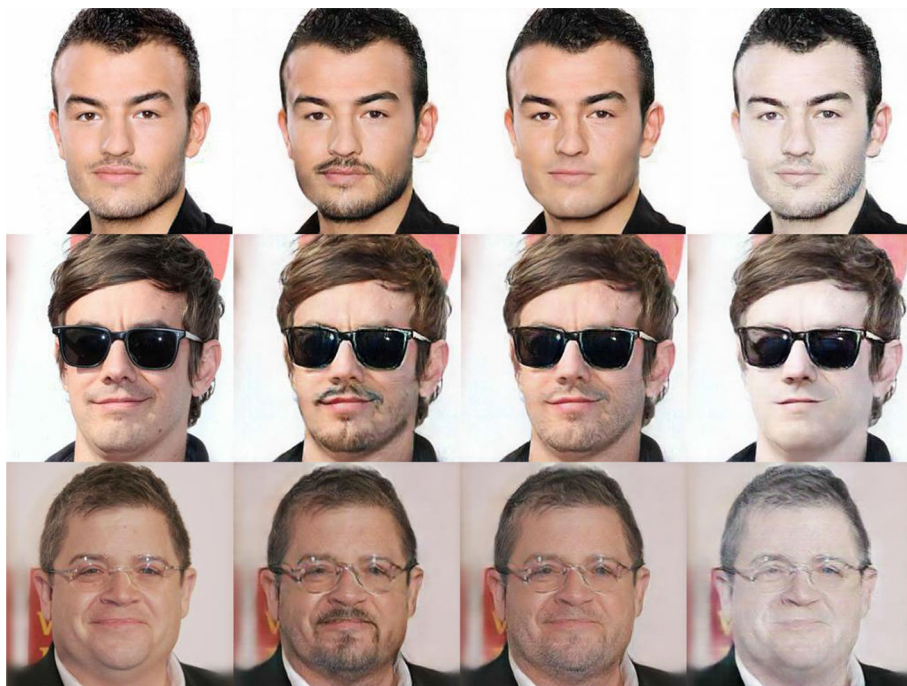
training. Conditional PixelCNN [136] can generate images conditioned on expression, pose and illumination. Furthermore, the works [46, 67, 74] transferred or swapped multiple attributes (hair color, bang, pose, gender, mouth open, etc.) among different faces simultaneously. Recently, Sanchez et al. [115] proposed a triple consistency loss to bridge the gap between the distributions of the input and generated images and allowed the generated images to be reintroduced to the network as input.

4 Transformation methods

In this section, we focus on the mapping function ϕ in Eq. 1 by reviewing the existing methods on face data augmentation. We divide these methods into basic image processing, model-based transformation, realism enhancement, generative-based transformation, augmented reality and autoaugmentation (Table 2) and give a clear description of them, respectively.

4.1 Basic image processing

The geometric and photometric transformations for generic data augmentation mainly utilize the traditional image processing algorithms. Digital image processing is an important research direction in the field of computer vision, which contains many simple and complex algorithms with a wide range of applications, such as classification, feature extraction, pattern recognition, etc. Digital image

transformation is a basic task of image processing, which can be expressed as:

$$g(x, y) = T[f(x, y)] \tag{3}$$

where $f(x, y)$ and $g(x, y)$ represent the input and output images, respectively, and T is the transformation function.

If T is an affine transformation, it could realize image translation, rotation, scaling, reflection, shearing, etc. Affine transformation is an important class of linear 2D geometric transformations which maps pixel intensity value located at position (x_1, y_1) in an input image to a new position (x_2, y_2) in an output image. The affine transformation is usually written in homogeneous coordinates as:

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & t_x \\ a_3 & a_4 & t_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \tag{4}$$

where (t_x, t_y) represents translation and parameters a_i contains rotation, scaling and shearing transformations.

If the image transformation is accomplished with a convolution between a kernel and an image, it can be used for image blurring, sharpening, embossing, edge enhancement, noise adding, etc. An image kernel k is a small two-dimensional matrix, and the image convolution can be expressed by the following equation:

Table 2 An overview of transformation methods

Methods	Implementations	Popular applications
Basic image processing	Noise adding, cropping, flipping, in-plane rotation, deformation, landmark perturbation, template fusion, mask blending, etc.	Generic transformation, expression synthesis, makeup transfer, hairstyle transfer, accessory wearing
Model-based transformation	e.g., 2D active appearance models	Pose transformation, expression transfer
	e.g., 3D morphable models	
Realism enhancement	Displacement map, augmentation function, detail refinement, domain adaption, etc.	Facial attributes transformation
Generative-based transformation	GANs, VAEs, PixelCNN, Glow, etc.	Facial attributes transformation
Augmented reality	Real and virtual fusion	Accessory wearing, hairstyle transfer
Autoaugmentation	Neural net, search algorithm, etc.	Generic transformation

$$g(x,y) = k(x,y)*f(x,y)$$
$$= \sum_{s=-a}^{a} \sum_{t=-b}^{b} k(s,t)f(x-s, y-t) \tag{5}$$

Meanwhile, the image color can also be linearly transformed through:

$$g(x,y) = w(x,y) \cdot f(x,y) + b \tag{6}$$

where $w(x,y)$ is the weight and b is the color value bias. This transformation can be performed in various color spaces including RGB, YUV, HSV, etc.

Some works performed a sequence of image manipulations to implement complex transformation for face data augmentation. For example, [94] perturbed the facial landmarks through a series of operations, including facial landmarks location, perturbation of the landmarks by Gaussian distribution and image normalization. [147] used face deformation and wrinkle mapping for facial expression synthesis. [49] proposed a pure 2D method to generate synthetic face images by compositing different face parts (eyes, nose, mouth) of two subjects. [8] synthesized new faces by stitching region-specific triangles from different images, which were proximal in a CNN feature representation space.

4.2 Model-based transformation

Model-based face data augmentation fits a face model to the input image and synthesizes faces with different appearance by varying the parameters of the fitted model. The commonly used generative face models can be classified as 2D and 3D, and the most representative models are 2D active appearance models (2D AAMs) [20] and 3D morphable models (3DMMs) [13]. Both the AAMs and 3DMMs consist of a linear shape model and a linear texture

model. The main difference between them is the shape component, which is 2D for AAM and 3D for 3DMM.

AAM is a parameterized face model represented as the variability of the face shape and texture. It is constructed based on a representative training set through a statistical-based template matching method and computed using principal component analysis (PCA). The shape of an AAM is described as a vector of coordinates from a series of landmark points and a landmark connectivity scheme. Mathematically, a shape of a 2D AAM is represented by concatenating n landmark points (x_i, y_i) into a vector $(x_1, y_1, x_2, y_2, \ldots, x_n, y_n)^T$.

With all the shape vectors from the training set, a mean shape is extracted. Then, the PCA is applied to search for directions that have large variance in the shape space and project the shape vectors onto it. Finally, the shape is modeled as a base shape plus a linear combination of shape variations:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{P}_s \cdot \mathbf{b}_s \tag{7}$$

In Eq. 7, $\bar{\mathbf{s}}$ denotes the base shape or mean shape, $\mathbf{P}_s$ is a set of orthogonal modes of variation obtained from the PCA eigenvectors. The coefficients $\mathbf{b}_s$ include the shape parameters in the shape subspace.

In AAM, the texture is defined with the pixel intensities. For t pixels sampled, the texture is expressed as $[g_1, g_2, \ldots, g_{ct}]^T$, where c is the number of channels in each pixel. After a piecewise affine warping and photometric normalization for all the texture vectors extracted from the training set, the texture model is obtained using PCA and has a similar expression with the shape model:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P}_g \cdot \mathbf{b}_g \tag{8}$$

where $\bar{\mathbf{g}}$ is the mean texture, $\mathbf{P}_g$ is the matrix consisting of a set of orthonormal base vectors, and $\mathbf{b}_g$ includes the texture parameters in the texture subspace. Figure 9 shows an
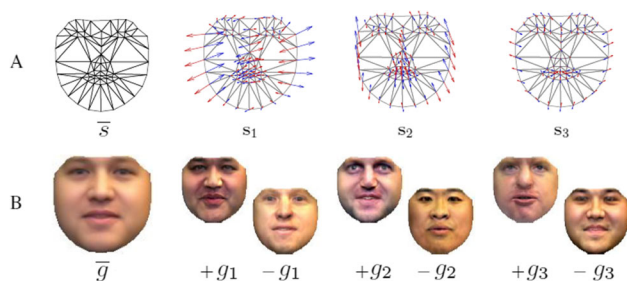
Fig. 9 Reprinted by permission from Springer Nature: Springer International Journal of Computer Vision [99], copyright 2007

example of 2D AAM constructed from 5 people using approximately 20 training images for each person. Figure 9a shows the mean shape $\bar{\mathbf{s}}$ and the first three shape variation components $\mathbf{s}_1$, $\mathbf{s}_2$ and $\mathbf{s}_3$. Figure 9b shows the mean texture $\bar{\mathbf{g}}$ and an illustration of the texture variation, where $+\mathbf{g}_j$ and $-\mathbf{g}_j$ denote the addition and subtraction of the jth texture mode to the mean texture, respectively.

3D morphable model is constructed from a set of 3D face scans with dense correspondence. The geometry of each scan is represented by a shape vector $\mathbf{S} = (x_1, y_1, z_1, \ldots, x_n, y_n, z_n)$, that contains the 3D coordinates of its n vertices, and the texture of the scan is represented by a texture vector $\mathbf{T} = (r_1, g_1, b_1, \ldots, r_n, g_n, b_n)$, that contains the color values of the n corresponding vertices. A morphable face model is constructed based on PCA and expressed as:

$$\mathbf{S}_{model} = \overline{\mathbf{S}} + \sum_{i=1}^{m} \alpha_i \mathbf{s}_i \quad \mathbf{T}_{model} = \overline{\mathbf{T}} + \sum_{i=1}^{m} \beta_i \mathbf{t}_i \qquad (9)$$

where m is the number of eigenvectors. $\overline{\mathbf{S}}$ and $\overline{\mathbf{T}}$ represent the mean shape and mean texture, respectively. $\mathbf{s}_i$ and $\mathbf{t}_i$ are the ith eigenvectors, and $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_m)$ and $\beta = (\beta_1, \beta_2, \ldots, \beta_m)$ are shape and texture parameters, respectively. 3DMM has very similar model expressions with 2D AAM. Figure 10 shows an example of 3DMM—SFM (Surrey Face Model) [55], which contains 3D meshes with multi-resolution. In this figure, we can visualize the mean

shape of SFM model along with the top five principal components of the shape variation.

To generate model instances (images), AAMs use a 2D image normalization and a 2D similarity transformation, while the 3DMMs use the scaled orthographic model or weak perspective model. More detailed description about the representational power, constriction and real-time fitting of the AAMs and 3DMMs can be found in the study of [99].

One advantage of 3D model is the availability of surface normals, which can be used to simulate the lighting effect. Therefore, some works combined the above 3DMMs with illumination modeled by Phong model [12] or Spherical Harmonic model [155] to generate more realistic face images. Hu et al. [50] classified the 3D face models into 3DSM (3D Shape Model), 3DMM and E-3DMM (Extended 3DMM). The 3DSM can only explicitly model pose. In contrast, 3DMM can model pose and illumination, while E-3DMM can model pose, illumination and facial expression. Furthermore, in order to overcome the limitation of linear models, Tran et al. [133] utilized neural networks to reconstruct nonlinear 3D face morphable model which is a more flexible representation. Hu et al. [50] proposed U-3DMM (Unified 3DMM) to model more intra-personal variations, such as occlusion.

Matching an AAM to an image can be considered as a registration problem and solved via energy optimization. In contrast, 3D model fitting is more complicated. The fitting is manly conducted by minimizing the color value differences over all the pixels in the facial region between the input images and its model-based reconstruction result. However, as the fitting is an ill-posed problem, it is not easy to get an efficient and accurate fitting [50]. Masi et al. [98] used the corresponding landmarks from the 2D input image and the 3D model to calculate the extrinsic camera parameters. Guo et al. [44] and Thies et al. [131] applied the analysis-by-synthesis strategy to do model fitting. In recent years, many works use neural networks to do 3D face model fitting [44, 165], and more fitting methods can be found in the study of [50].

Although 3D model can be used to generate more diverse and accurate transformations of faces, there still exist some challenges. One of the challenges is the visualization of teeth and mouth cavity, as the face model only represents the skin surface and does not include eyes, teeth and mouth cavity. Thies et al. [131] used two textured 3D proxies for the teeth simulation and achieved mouth transformation by wrapping a static frame of an open mouth based on the tracked landmarks. Another challenge is the artifacts caused by the missing of occluded regions when the head pose is changed. Zhu et al. [166] proposed an inpainting method which made use of Poisson editing to
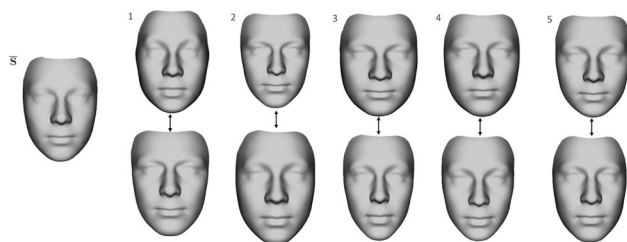

Fig. 10 Visualization of the face shape of SFM [55]. The leftmost shows the visualization of the mean shape. The others are the visualizations of the first five principal components of shape, with each visualized as additions and subtractions from the mean

estimate the mean face texture and fill the facial detail of the invisible region caused by self-occlusion.

4.3 Realism enhancement

As illustrated in Fig. 1, besides the direct style transfer from real 2D images, making simulated samples more realistic is an important method of face data augmentation. Although modern computer graphics techniques provide powerful tools to generate virtual faces, it still remains difficult to generate a large number of photorealistic samples due to the lack of accurate illumination and complicated surface modeling. The state of the art synthesizes virtual faces using a morphable model and has difficulty in generating detailed photorealistic images of faces, such as faces with wrinkles [44]. Anyway, the simulation process involved is simplified for the consideration of the speed of modeling and rendering, making the generated images not realistic enough. In order to improve the quality of simulated samples, some realism enhancement techniques were proposed.

Guo et al. [44] introduced displacement map that encoded the geometry details of face in a displacement along the depth direction of each pixel. It could be used for face detail transfer and synthesizing fine-detailed faces. Zhao et al. [160] applied GANs to improve the realism of face simulator's output by making use of unlabeled real faces. Specifically, they fed the defective simulated faces obtained from a 3D morphable model into a generator for realism refinement and used two discriminators to minimize the gap between real domain and virtual domain by discriminating real *vs* fake and preserving identity information. Furthermore, Gecer et al. [36] introduced a two-way domain adaption framework similar to CycleGAN to improve the realism of rendered faces. Zhao et al. [158] and Huang et al. [54] applied dual-path GANs, which contained separate global generator and local generators for global structure and local details generation. Shrivastava et al. [120] proposed SimGAN, a simulated and unsupervised learning method to improve the realism of synthetic images using a refiner network and adversarial training. Sixt et al. [124] embedded a simple 3D model and a series of parameterized augmentation functions into the generator network of GAN, where the 3D model was used to produce virtual samples from input labels and the augmentation functions used to add the missing characteristics to the model output. Although the proposed realism enhancement methods proposed by Shrivastava et al. [120] and Sixt et al. [124] were not originally aimed at face images, they have the potential to be applied to realistic face data generation.

4.4 Generative-based transformation

The generative models provide a powerful tool to generate new data from modeled distribution by learning the data distribution of the training set. Mathematically, the generative model can be expressed as follows. Suppose there is a dataset of examples $\{x_1, \ldots, x_n\}$ as samples from a real data distribution $p(x)$ as illustrated in Fig. 11, in which the green region shows a subspace of the image space containing real images. The generative model maps a unit Gaussian distribution (gray) to another distribution $\hat{p}(x)$ (blue) through a neural network, which is a function with parameters $\boldsymbol{\theta}$. The generated distribution of images can be tweaked when the network parameters are changed. Then the aim of the training process is to start from random and find parameters $\boldsymbol{\theta}$ that produce a distribution that closely matches the real data distribution.

In recent years, with the rapid growing of computational resources, the deep generative models have attracted much attention and significantly promoted the performance of data generation. Among them, the three most popular models are autoregressive models, variational autoencoders (VAEs) and generative adversarial networks. Autoregressive models and VAEs aim to minimize the Kullback–Leibler (KL) divergence between the modeled distribution and the real data distribution. In contrast, the generative adversarial networks apply adversarial learning to generate data indistinguishable from the real samples and hence avoid specifying an explicit density for any data point, which belong to the class of implicit generative models [38].

4.4.1 Autoregressive generative models

The typical examples of autoregressive generative models are PixelRNN and PixelCNN proposed by Van Den Oord et al. [137]. They tractably model the joint distribution of the pixels in the image by decomposing it into a product of conditional distributions, which can be formulated as:

$$p(\mathbf{x}) = \prod_{i=1}^{n^2} p(x_i|x_1, \ldots, x_{i-1}) \tag{10}$$

where $p(\mathbf{x})$ is the probability of image $\mathbf{x}$ formed of $n \times n$ pixels and the value $p(x_i|x_1, \ldots, x_{i-1})$ is the probability of the ith pixel x_i given all the previous pixels $x_1, \ldots, x_{i-1}$. Thus, the image modeling problem turns into a sequential problem, where one learns to predict the next pixel given all the previously generated pixels. The PixelRNN models the pixel distribution with two-dimensional LSTM and PixelCNN models with convolutional networks. Van den Oord et al. [136] further presented Gated PixelCNN and

Fig. 11 A schematic diagram of generative model. The unit Gaussian distribution is mapped to a generated data distribution by the generative model, and the distance between the generated data distribution and the real data distribution is measured by the loss
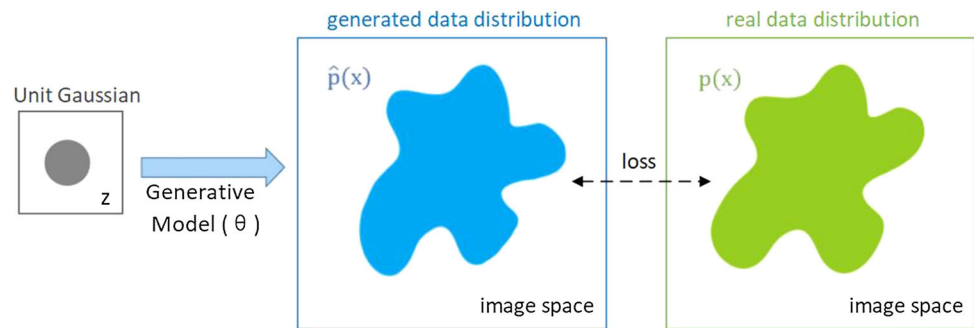
Conditional PixelCNN. The former replaced the activation unit in the original PixelCNN with gated block, and the latter modeled the complex conditional distributions of natural images by introducing conditional variant to the latent vector. In addition, Salimans et al. [114] proposed PixelCNN++ which simplified PixelCNN's structure and improved the synthetic images' quality.

4.4.2 Variational autoencoders

Variational autoencoders formalize the data generation problem in the framework of probabilistic graphical models rooted in Bayesian inference. The idea of VAEs is to learn the latent variables, which are low-dimensional latent representations of the training data and inferred through a probabilistic model. In Fig. 12-left, the latent variables are denoted by $\mathbf{z}$, and the probability distribution of $\mathbf{z}$ is denoted as $p_\theta(\mathbf{z})$, where θ are the model parameters. There are two components in a VAE: the encoder and the decoder. The encoder encodes the training data $\mathbf{x}$ into a latent representation, and the decoder maps the obtained latent representation $\mathbf{z}$ back to the data space. In order to estimate the model parameters θ, we maximize the likelihood of the training data:

$$p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}, \mathbf{z}) \, d\mathbf{z} = \int p_\theta(\mathbf{x}|\mathbf{z}) p_\theta(\mathbf{z}) \, d\mathbf{z} \qquad (11)$$

The above integral is intractable, and the true posterior density $p_\theta(\mathbf{z}|\mathbf{x}) = p_\theta(\mathbf{x}|\mathbf{z})p_\theta(\mathbf{z})/p_\theta(\mathbf{x})$ is intractable, so the EM (Expectation–Maximization) algorithm cannot be used. The required integrals for any reasonable mean-field

VB (variational Bayesian) algorithm are also intractable [69]. Therefore, the VAEs turn to infer $p(\mathbf{z}|\mathbf{x})$ using variational inference which is a basic optimization problem in Bayesian statistics. They first model $p(\mathbf{z}|\mathbf{x})$ using simpler distribution $q_\phi(\mathbf{z}|\mathbf{x})$ which is easy to find and try to minimize the difference between $p_\theta(\mathbf{z}|\mathbf{x})$ and $q_\phi(\mathbf{z}|\mathbf{x})$ using KL divergence metric approach. The marginal likelihood of individual datapoint can be written as:

$$\log p_\theta(\mathbf{x}) = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) + \mathcal{L}(\theta, \phi; \mathbf{x}) \qquad (12)$$

where the first term of the right-hand side is the KL divergence of the approximate from the true posterior. Since this divergence is non-negative, the second term $\mathcal{L}(\theta, \phi; x)$ represents the lower bound on the marginal likelihood of the datapoint which we want to optimize and can be written as:

$$\mathcal{L}(\theta, \phi; \mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) \qquad (13)$$

More detailed explanation of the mathematical derivation can be found in [69].

In order to control the generation direction of the VAEs, Sohn et al. [125] proposed a conditional variational autoencoder (CVAE), which is a conditional directed graphical model being trained to maximize the conditional log-likelihood. Pandey et al. [106] introduced conditional multimodel autoencoder (CMMA) to address the problem of conditional modality learning through the capture of conditional distribution. Their model can be used to generate and modify faces conditioned on facial attributes.

Fig. 12 Illustration of variational autoencoders and generative adversarial networks
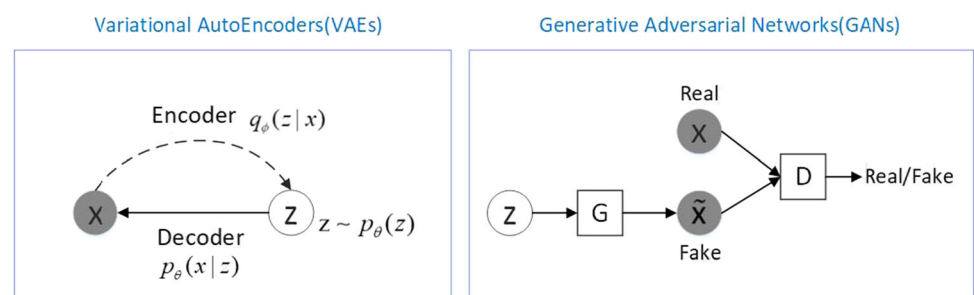
Another approach to generate faces from visual attributes was introduced by Yan et al. [148]. They modeled the image as a composition of foreground and background and developed a layered generative model with disentangled latent variables that were learned using a variational autoencoder. Huang et al. [52] proposed an introspective variational autoencoder (IntroVAE) model for synthesizing high-resolution photorealistic images by self-evaluating the quality of the generated samples during the training process. They borrowed the idea of GANs and reused the encoder as a discriminator to classify the generated and training samples.

4.4.3 Generative adversarial network

Generative adversarial network is an alternative framework to train generative models which gets rid of the difficult approximation of intractable probabilistic computations. It takes game-theoretic approach and plays an adversarial game between a generator and a discriminator. The discriminator learns to distinguish between real and fake samples, while the generator learns to produce fake samples that are indistinguishable from real samples by the discriminator.

As shown in Fig. 12-right, in order to learn a generated distribution p_g over data $\mathbf{x}$, the generator builds a mapping function from a prior noise distribution $p_{\mathbf{z}}(\mathbf{z})$ to a data space as $G(\mathbf{z};\boldsymbol{\theta}_g)$, where G is a differentiable function represented by a multilayer perceptron with parameters $\boldsymbol{\theta}_g$. The discriminator, whose mapping function is denoted as $D(\mathbf{x};\boldsymbol{\theta}_d)$, outputs a single scalar representing the probability that $\mathbf{x}$ comes from the training data rather than p_g. G and D are trained simultaneously by adjusting the parameters of G to minimize $\log\left(1 - D(G(\mathbf{z}))\right)$ and also the parameters of D to maximize the probability of assigning the correct label to both training examples and generated samples. The objective can be represented by the following two-player min–max game with value function $V(D, G)$:

$$\min_G \max_D V(D, G) = \mathbb{E}_{\mathbf{x}\sim\, p_{data}(\mathbf{x})}[\log D(\mathbf{x})]$$
$$+ \mathbb{E}_{\mathbf{z}\sim\, p_{\mathbf{z}}(\mathbf{z})}[\log\left(1 - D(G(\mathbf{z}))\right)] \quad (14)$$

Ever since the proposal of the basic GAN concept, researchers have been trying to improve its stability and capability. Radford et al. [112] introduced DCGANs, the deep convolutional generative adversarial networks, to replace the multilayer perceptrons in the basic GANs with convolutional nets. Whereafter, a lot of improvements for DCGANs were proposed. For example, [113] improved the training techniques to encourage convergence of the GANs game, [6] applied Wasserstein distance and improved the stability of learning, Conditional GANs could determine the specific representation of the generated images by feeding the condition to both the generator and discriminator [56, 101]. What's more, Zhang et al. [154] proposed a two-stage GANs–StackGAN. The Stage-I GAN sketched the primitive shape and colors of the object, and the Stage-II GAN generated realistic high-resolution images based on the Stage-I's output. Chen et al. [18] introduced InfoGAN, an information-theoretic extension to the basic GANs. It used a part of the input noise vector as latent code to target the salient structured semantic features of the data distribution in an unsupervised way.

Since the first proposition of GANs [37] in 2014, it has attracted much attention because of its remarkable performance in a wide range of applications, including face data augmentation. Antoniou et al. [5] proposed Data Augmentation Generative Adversarial Network (DA-GAN) based on conditional GAN (cGAN) and tested its effectiveness on vanilla classifiers and one shot learning. Actually, it is a basic framework for data augmentation using cGAN. Many face data augmentation works followed this architecture and extended it to a more powerful network. We redraw this basic architecture in Fig. 13, which comprises a conditional image generation and discrimination network.

Zhu et al. [167] presented another basic framework for face data augmentation based on CycleGAN [164]. Similar to cGAN, CycleGAN is also a general-purpose solution for image-to-image translation, but it learns a dual mapping between two domains simultaneously with no need for paired training examples, because it combines a cycle consistency loss with adversarial loss. The authors in [167] used this framework (whose architecture is shown in Fig. 14) to generate auxiliary data for unbalanced dataset, where the data class with fewer samples was selected as transfer target and the data class with more samples was reference. In Fig. 14, both the reference images and target images flow into the CycleGAN as domains R and T. G and F are two generators to transfer $R \rightarrow T$ and $T \rightarrow R$. Two discriminators are represented by $D(R)$ and $D(T)$, respectively, where $D(R)$ aims to distinguish between the real images in R and the generated fake images in $F(T)$, and $D(T)$ aims to distinguish between the real images in T and the generated fake images in $G(R)$. What's more, the cycle consistency loss was used to guarantee that $F(G(R)) \approx R$ and $G(F(T)) \approx T$. Zhu et al. [167] made a comparison between CycleGAN and the classical GAN. They claimed that the original GAN learns a mapping from low-dimensional manifold (determined by noise) to high-dimensional data spaces (images), while CycleGAN learns the translation between two high-dimensional data domains, which means CycleGAN can complete and complement an imbalanced dataset more efficiently.
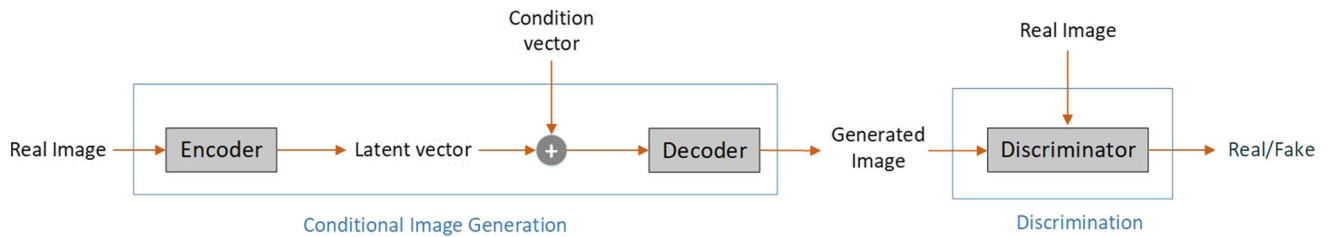
Fig. 13 Face data augmentation architecture based on cGAN. Generation: An encoder maps the input image into a lower-dimensional latent space while a condition vector is transformed and concatenated with the latent vector. Then the long vector is passed to the decoder to generate an augmentation image. Discrimination: In order to ensure the realism of the generated image, an adversarial discriminator network is employed to discriminate the generated images from the real images

Fig. 14 Reprinted by permission from Springer Nature: Springer eBook [167], copyright 2018
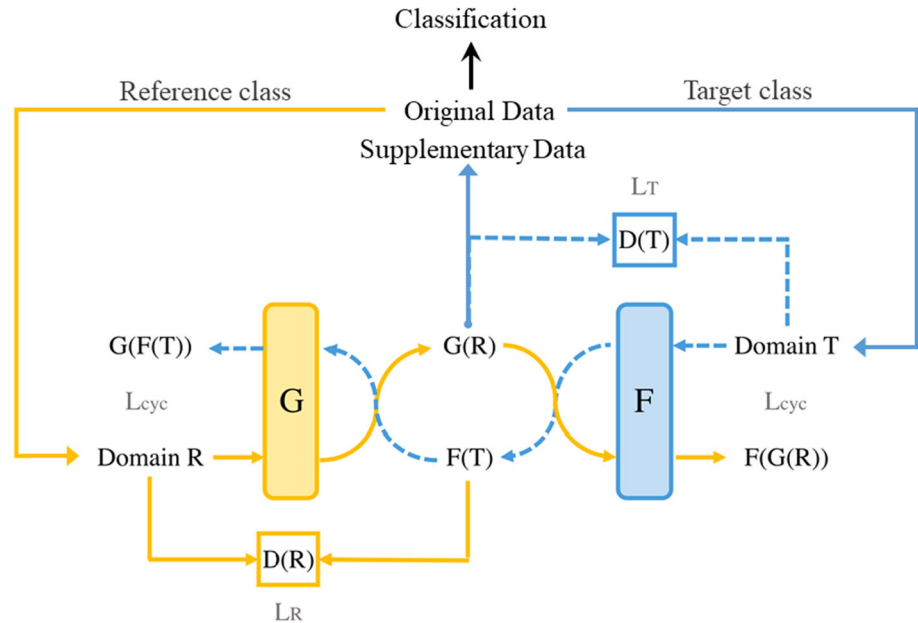
Except the above two basic frameworks of face data augmentation based on GAN, numerous extended approaches were proposed in recent years, such as DiscoGAN [66], StarGAN [19], F-GAN [105], Age-cGAN [4], IPC-GANs [141], BeautyGAN [82], PairedCycleGAN [16], G2-GAN [127], GANimation [110], GC-GAN [111], ExprGAN [29], DR-GAN [134], CAPG-GAN [51], UV-GAN [27], CVAE-GAN [10], RenderGAN [124], DA-GAN [159, 160], TP-GAN [54], SimGAN [120], FF-GAN [151], GP-GAN [28], Attribute-GAN [86, 153], PGGAN [61], and so on. They all greatly promoted the application of GANs, especially the proposition of high-resolution face image generation. Karras et al. [61] introduced an approach to generate large high-quality images by progressively adding layers to the generator and discriminator model during the training process. Moreover, it is more stable and efficient.

4.4.4 Flow-based generative models

In addition to autoregressive models and VAEs, flow-based generative models are likelihood-based generative methods as well, which were first described in NICE [30]. In order to model complex high-dimensional densities, they first map the data to a latent space where the distribution is easy to model. The mapping is performed through a nonlinear deterministic transformation which can be decomposed into a sequence of transformations and is invertible. Let $\mathbf{x}$ be a high-dimensional random vector with unknown distribution, $\mathbf{z}$ be the latent variable, the relationship between the $\mathbf{x}$ and $\mathbf{z}$ can be written as:

$$\mathbf{x} \xleftrightarrow{f_1} h_1 \xleftrightarrow{f_2} h_2 \cdots \xleftrightarrow{f_k} \mathbf{z} \tag{15}$$

where $\mathbf{z} = f(\mathbf{x})$ and $f = f_1 \circ f_2 \circ \cdots \circ f_k$. Such a sequence of invertible transformations is called a flow.

Flow-based generative models have not gained much attention so far. In fact, they have exact latent-variable inference and log-likelihood evaluation comparing to

VAEs and GANs, and they are able to perform efficient inference and synthesis comparing to autoregressive models [68]. Kingma and Dhariwal [68] proposed Glow, a simple type of generative flow using an invertible 1×1 convolution. They demonstrated the model's ability in synthesizing high-resolution face images through a series of experiments for image generation, interpolation and semantic manipulation. Although the results still had a gap with GANs, they showed significant improvement against previous flow-based generative models. Grover et al. [38] introduced Flow-GAN, a generative adversarial network with a normalizing flow generator, with the purpose of bridging the gap between high-quality generated samples and ill-defined likelihood for GANs. It transformed the prior noise density into a model density through a sequence of invertible transformations, so the exact likelihoods could be tractably evaluated.

4.4.5 Generative models comparison

Each generative model has its pros and cons. Accurate likelihood evaluation and sampling are tractable in autoregressive models. They have given the best log likelihoods so far and have stable training process. However, they are less effective during sampling and the sequential generation process is slow. Variational autoencoders allow us to perform both learning and efficient inference in sophisticated probabilistic graphical models with approximate latent variables. Anyway, the likelihood is intractable to compute, and the variational lower bound to optimize for learning the model is not as exact as that in autoregressive models. Meanwhile, their generated samples tend to be blurry and of lower quality compared to GANs. GANs can generate sharp images, and there is no Markov chain or approx networks involved during sampling. However, they cannot provide explicit density. This makes it more challenging for quantitative evaluations [38] and also more difficult to optimize due to unstable training dynamics.

In order to improve the performance of generative models, many efforts have been made. For example, some works modified the architecture of these models to gain better characters, such as Gated PixelCNN [136], CVAE [125] and DCGAN [112]. Some works try to combine different models in one generative framework. An example of the combination of VAEs and PixelCNN is PixelVAE [41]. It is a VAE model with an autoregressive decoder based on PixelCNN, whereas it has fewer autoregressive layers than PixelCNN and learns more compressed latent representations than standard VAE. Perarnau et al. [108] combined an encoder with a cGAN into IcGAN (Invertible cGAN), which enabled image generation with deterministic modification. [75, 95] proposed similar ideas by

combining VAE with GANs and introduced AAE (adversarial autoencoder) and VAE/GAN, respectively, which could generate photorealistic images while keeping training stale. What's more, Zhang et al. [156] designed CAAE (conditional adversarial autoencoder) for face age progression and regression. Zhou et al. [162] presented CDAAE (conditional difference adversarial autoencoder) for facial expression synthesis. Bao et al. [10] proposed CVAE-GAN, a conditional variational generative adversarial network capable of generating images of fine-grained object categories, such as faces of a specific person.

4.5 Augmented reality

Augmented reality (AR) is a technique which supplements the real word with virtual (computer-generated) objects that appear to coexist in the same space [135]. It allows a seamless fusion of virtual elements and real world, like fusing a real desk with a virtual lamp, or trying virtual glasses on a real face (Fig. 15). The application of AR technology can expand the scale of training data from the following aspects: supplementing the missing elements in the real scene, providing precise and detailed virtual elements and increasing the data diversity.

In order to improve the performance of eyeglasses face recognition, Guo et al. [43] synthesized such images by reconstructing 3D face models and fitting 3D eyeglasses based on anchor points. The experiments on the real face dataset validated that their synthesized data had expected improvement in face recognition. LiteOn also generated augmented face images with eyeglasses to increase the accuracy of face recognition [143]. In fact, many AR techniques can be adopted for face data augmentation, such as the virtual mirror [70] proposed for facial geometric alteration, the magic mirror [57] used for makeup or
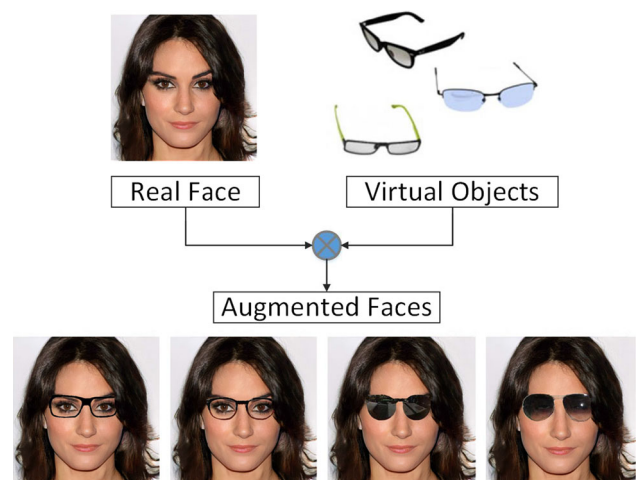

Fig. 15 Applying AR for data augmentation. The real face is fused with virtual glasses and turns into an augmented face wearing glasses

accessories try-on, the Beauty e-Experts system [85] designed for hairstyle and facial makeup recommendation and synthesis, the virtual glasses try-on system [7], etc. More experiences of data augmentation by AR can be found in the studies of [3, 31, 100].

4.6 Autoaugmentation

Different augmentation strategies should be applied to different tasks. In order to automatically find the most appropriate scheme to enrich the training set and obtain an optimal result, some autoaugmentation methods were proposed. For example, Wang and Perez [139] applied an Augmentation Network (AugNet) to learn the best augmentation for improving a classifier, Lemley et al. [77] presented Smart Augmentation by creating one or multiple networks to learn the suitable augmentation for the given class of input data during the training of the target network. Different from the above works, Cubuk et al. [22] created a search space for data augmentation policies and used reinforcement learning as the search algorithm to find the best operations of data augmentation on the dataset of interest. During the training of the target network, it used a controller to predict augmentation decisions and evaluated them on a validation set in order to produce reward signals for the training of the controller. Remarkably, the current autoaugmentation techniques are usually designed for simple augmentation operations, such as rotation, translation and cropping.

5 Evaluation metrics

The two main methods of evaluation are the qualitative evaluation and quantitative evaluation. For qualitative evaluation, the authors directly present the readers or interviewers with generated images and let them judge the quality. The quantitative evaluation is usually based on some statistical methods. It provides quantifiable and explicit results. Most of the time, both qualitative and quantitative methods are employed together to provide adequate information for the evaluation, and suitable evaluation metrics should be applied since face data augmentation includes different transformation types with different purpose.

The frequently used metrics include the accuracy and error rate, distance measurement, Inception Score (IS) [113] and Fréchet Inception Distance (FID) [47], which are introduced, respectively, as follows.

Accuracy and error rate are the most commonly used measurements for classification and prediction, which are calculated on the numbers of positive and negative samples. Assume the numbers of positive samples and negative samples are N_p and N_n, and the numbers of correct predictions for the positive and negative are t_p and t_n, respectively. The accuracy rate is defined as $AccuracyRate = ((t_p + t_n)/(N_p + N_n))$. The error rate is $ErrorRate = 1 - AccuracyRate$. However, the above equations only work for balanced data, which means the positive samples have equal number with the negative. Otherwise, it should adopt the balanced accuracy and error rate, which are defined as $BalancedAccuracyRate = 1/2(t_p/N_p + t_n/N_n)$, and $BalabcedErrorRate = 1 - BalancedAccuracyRate$.

Distance measurement can be used in a wide range of scenarios. For example, the L1 norm and L2 norm are usually adopted to calculate the color distance and spatial distance, the peak signal-to-noise ratio (PSNR) is used to measure pixel differences, and the distance of two probability distributions can be measured by the KL divergence or Fréchet distance. Usually, the distance metrics are used as the basis of other metrics.

Inception Score measures the performance of a generative model from the quality of the generated images and their diversity. It is computed based on the relative entropy of two probability distributions which are relative to the output of a pretrained classification network. The IS is defined as $\exp(\mathbb{E}_{x \sim p_g} D_{KL}(p(y|x)||p(y)))$, where x denotes the generated image, y is the class label output by the pretrained network and p_g represents the generated distribution. The principle of IS is that the images generated by a better generative model should have a conditional label distribution $p(y|x)$ with low entropy (which means the generated images are highly predictable) and a marginal distribution $p(y) = \int p(y|x = G(z))dz$ with high entropy (which means the generated images are diverse).

Fréchet Inception Distance applies an inception network to extract image feature and models the feature distributions of the generated and real images as two multivariate Gaussian distributions. The quality and diversity of the generated images are measured by calculating the Fréchet distance of the two distributions, which can be expressed as $\text{FID}(x, g) = ||\mu_x - \mu_g||_2^2 + \text{Tr}(\sum_x + \sum_g - 2(\sum_x \sum_g)^{\frac{1}{2}})$, where $(\mu_x, \sum_x)$ and $(\mu_g, \sum_g)$ are the mean and covariance of the two Gaussian distributions correlated with the real images and the generated images. In comparison with IS, FID is more robust to noise and more sensitive to intra-class mode collapse [53].

Actually, it is difficult to give a totally fair comparison of different face augmentation methods with uniform criteria, as they usually focus on diverse problems and are designed for different applications. For example, the image quality and diversity are the main concerns for image generation, so Inception Score and FID are the most widely adopted metrics, whereas, for conditional image

generation, the generation direction or the generated images' domain is also important. Furthermore, if the work focuses on identity preservation, a face recognition or verification test is necessary. Usually, the importance of each component inside the framework is evaluated through ablation study. If the effect of a data augmentation method on a specific task is desired to be evaluated, a direct way is to compare the task execution result with and without the augmented data. One notable thing is that the task execution is based on specific algorithms and datasets, so it is impossible to evaluate over all possible algorithms and datasets.

6 Challenges and opportunities

Face data augmentation is effective on enlarging the limited training dataset and improving the performance of face-related detection and recognition tasks. Despite the promising performance achieved by massive existing works, there are still some issues to be tackled. In this section, we point out some challenges and interesting directions for the future research of face data augmentation.

6.1 Identity preserving

Although many facial attributes are transformed during the data augmentation procedure, the identity which is the most important class label for face recognition and verification is usually expected to preserve in most cases. However, it still remains difficult in conditional face generation [119].

So far, the most commonly used approach in existing works is the introduction of an identity loss during network training. For example, Cao et al. [15] added a perceptual identity loss to integrate domain knowledge of identity into the network, which is similar to [54, 80, 127, 141]. They employed the method proposed in the study of [58] to calculate the perceptual loss and extracted identity features through a face recognition module, e.g., VGG-Face [107] or Light CNN [145]. In addition, the works [4, 11, 119, 142] limited the identity feature distance (Manhattan distance or Euclidean distance) to decrease the identity loss in the process of face image generation. Deng et al. [27] defined a center loss for identity, based on the activations after the average pooling layer of ResNet.

Another way of identity preservation is to use cycle consistency loss to supervise the identity variation. For example, PairedCycleGAN [16] contained an asymmetric makeup transfer framework for makeup applying and removal, where the output of the two asymmetric functions should get back the input. In the study of [105], the cycle

consistency loss for identity preserving is defined by comparing the original face with the aging-and-rejuvenating result. Some work applied both the perceptual loss and cycle consistency loss for facial identity preservation [82].

In addition to the methods mentioned above, some researchers adopted specific networks to supervise identity changes. Tran et al. [134] modified the classical discriminator of GANs. In their DR-GAN for face pose synthesizing, the discriminator was trained to not only distinguish real and synthetic images, but also predict the identity and pose of a face. However, their multi-task discriminator classified all the synthesized faces as one class. In contrast, Shen et al. [118] proposed a three-player GAN, where the face classifier was treated as the third layer and competed with the generator. Moreover, their classifier differentiated the real and synthesized domains by assigning them different identity labels. The author claimed that previous methods cannot satisfy the requirement of identity preservation because they only tried to push real and synthesized domains close, but neglected how close they were.

Despite the efforts have been made, identity preservation is still a challenging problem which has not been completely solved. In fact, identity is presented by various facial attributes. Therefore, it is important to maintain the identity-relative features when changing other attributes, while this remains difficult for an end-to-end network.

6.2 Disentangled representation

Disentangled representation can improve the performance of neural networks in conditional generation by adjusting corresponding factors while keeping other attributes fixed. It enables a better control of the network output. For this purpose, Shu et al. [123] introduced a special generative adversarial network by encoding the image formation and shading processes into network layers. In consequence, the network could infer a face-specific disentangled representation of intrinsic face properties (shape, albedo and lighting) and allowed for semantically relevant facial editing. Qiao et al. [111] used separate paths to learn the geometry expression feature and image identity feature in a disentangled manner. Bao et al. [11] disentangled the identity and other attributes of faces by introducing an identity network and an attribute network to encode the identity and attributes into separate vectors before importing them to the generator. Shen et al. [119] proposed a two-stage approach for face synthesis. It produced disentangled facial features from random noises using a series of feature generators in the first stage and decoded these features into synthetic images through an image generator in the second stage. Chen et al. [17] disentangled the texture and deformation of the input images by adopting the Intrinsic DAE (Deforming Autoencoder)

model [122]. It transferred the input image to three physical image signals (shading, albedo and deformation) and combined the shading and albedo to generate texture image. Liu et al. [84] introduced a composite 3D face shape model composed of mean face shape, identity-sensitive difference and identity-irrelevant difference. They disentangled the identity and non-identity features in the 3D face shape and represented them with separate latent variables.

The encoder–decoder architecture is widely used for face editing by mapping the input image into a latent representation and reconstructing a new face with desired attribute. Tran et al. [134] disentangled the pose variation from identity representation by inputting a separate pose code to the decoder and validating the generated pose with the discriminator. Chen et al. [18] learned disentangled and interpretable representations for images in an entirely unsupervised manner by adding new objective to maximize the mutual information between small subsets of the latent variables and the observation. Zhou et al. [161] proposed GeneGAN, whose encoder decomposed the input image to background feature and object feature. Xiao et al. [146] constructed a DNA-like latent representation, in which different pieces of encodings controlled different visual attributes. Similarly, Lample et al. [74] tried to learn a disentangled latent space for explicit attribute control by applying adversarial training in latent space instead of the pixel space. However, as argued by He et al. [46], the attribute-independent constraint on the latent representation was excessive, as it may restrict the capacity of the latent representation and cause information loss. Instead, they defined constraint for the generated images through an attribute classification for accurate facial attribute editing.

6.3 Unsupervised data augmentation

Collecting large amounts of images with certain attributes is a difficult task, which limits the application of data augmentation based on supervised learning. Therefore, semi-supervised and unsupervised methods are proposed to reduce the data demand of face generation. Wiles et al. [142] produced generated faces with the pose and expression of the driving frame without requiring expression or pose label, or coupled training data either. In the training stage, it extracted source frame and driving frame from a same video, so the generated and driving frames would match. Kim et al. [66] implemented cross-domain translation without any explicitly paired data. In the study of [156], only multiple faces with different ages were used, and no paired samples for training or labeled faces for test were needed. Song et al. [126] developed a dual conditional GANs (Dual cGANs) for face aging and rejuvenation without the requirement of sequential training data. Shen

and Liu [117] adopted dual learning, which could transform images inversely and learn from each other. Moniz et al. [102] inferred the depth of facial keypoints without using any ground-truth of depth information. Kim et al. [67] and Pumarola et al. [110] proposed unsupervised strategies for visual attributes transfer and expression transfer. Zhu et al. [164] presented cycleGAN, which could be used in image-to-image translation with no need for paired training data. Then Song et al. [127] employed cycleGAN to simultaneously perform expression generation and removal. Chang et al. [16] applied cycleGAN for makeup applying and removal. Recently, Lu et al. [92] proposed conditional CycleGAN for conditional face generation, which was also an unsupervised learning method.

Both the supervised and unsupervised learning methods have their respective pros and cons. On the one hand, unsupervised learning methods make the preparing of training data much easier. It has boosted the development and application of face data augmentation. On the other hand, without the help of appropriately classified and labeled data, the learning process becomes more difficult and unstable, and the learned model is less accurate. In the study of [164], a lingering gap between the results achievable with paired training data and unpaired data was observed, and it was believed that this gap would be hard or even impossible to close in some cases. In order to make up the defects in training data, extra information should be injected, such as prior knowledge or expert knowledge. Actually, image translation in unsupervised setting is a highly ill-posed problem, since there exist an infinite set of joint distributions that can achieve the two domains translation [87]. Therefore, more effort should be made to lower the training difficulty if less training data is desired.

6.4 Improvement of GANs

In recent years, GANs have become one of the most popular methods in data generation. It is easy to use and has created a lot of impressive results. However, it still suffers from problems like instable training and mode collapse. Efforts to improve the effectiveness of GANs have never been stopped. Besides the works introduced in Sect. 4.4.3, many researchers modified the loss functions for higher quality of the generated results. For example, [82] applied four types of losses, including the adversarial loss, cycle consistency loss, perceptual loss and makeup constrain loss, to guarantee the quality of the makeup transfer. Choi et al. [19] used adversarial loss, domain loss and reconstruction loss for the training of their StarGAN. As mentioned in Sect. 6.1, many works adopted identity loss to preserve the identity of the generated face.

Recently, some improvement for the architecture of the original GANs were presented. Karras et al. [61] proposed a new training methodology for GANs to generate high-quality images. They started the training process from low resolution and progressively add layers to the generator and discriminator for a higher resolution. Shen et al. [118] proposed a symmetry three-player GAN—FaceID-GAN, in contrast to the classical two-player game of most GANs. It applied a face identity classifier as the third player to distinguish the identities of the real and synthesized faces and collaborate with the discriminator to compete generator. Karras et al. [62] redesigned the generator architecture of GANs and proposed a style-based generator. They mapped the input latent code to an intermediate latent space and employed Gaussian noise to create stochastic and localized variation to the generated images. Inspired by the natural images that exhibit multi-scale characteristics along the hierarchical architecture, Yang et al. [149] proposed a pyramid architecture for the discriminator of GANs, which estimated the pyramid facial feature representations at multiple scales and handled face generation in a fine-gained way. Their evaluation result demonstrated that the pyramid structure advanced the generation of aged faces by making them more natural and possessing more face details. Banerjee et al. [8] proposed MAD-GAN (Multi-Agent Diverse GAN), which incorporated multiple generators and one discriminator. Gu et al. [39] proposed a new differential discriminator and a network architecture, Differential Generative Adversarial Networks (D-GAN), in order to approximate the face manifold for nonlinear facial variations with small amount of training data. Kossaifi et al. [72] presented a method to incorporate geometry information into face image generation and introduced the Geometry-Aware Generative Adversarial Networks (GAGAN). Juefei-Xu et al. [59] introduced a stage-wise learning paradigm for GANs that ranked multiple stages of generators by comparing the margin-based ranking loss of the generated samples. Tian et al. [132] proposed a two-pathway (generation path + reconstruction path) framework, CR-GAN, to develop the widely used single-pathway (encoder–decoder–discriminator) network. The two pathways combined with self-supervised learning can learn complete representations in the embedding space and produce high-quality image generations from unseen data in wild conditions.

7 Discussion

The lack of labeled samples has been a common problem for researchers and engineers working with deep learning. Undoubtedly, data augmentation is an effective tool for solving this problem and has been widely used in various tasks. Among these tasks, face data augmentation is more complicated and challenging than others. Various methods have been proposed to transform a real face image to a new type, such as pose transfer, hairstyle transfer, expression transfer, makeup transfer and age transfer. Meanwhile, the simulated virtual faces can also be enhanced to be as realistic as the real ones. All these augmentation methods can be used to increase the variation of the training data and improve the robustness of the learned model.

Image processing is a traditional but powerful method for image transformation, which has been widely used in geometric and photometric transformations of generic data. It is more suitable for transforming the entire image in a uniform manner, rather than changing the facial attributes that need to transform a specific part or property of faces. Model-based method is intuitively suitable for virtual face generation. With the reconstructed face model, it is easy to modify the shape, texture, illumination and expression. However, it remains difficult to reconstruct a complete and precise face model from a single 2D image, and synthesizing a virtual face to be really realistic is still computationally expensive. Therefore, several realism enhancement algorithms were proposed. With the rapid development of deep learning, learning-based method has become more and more popular. Although challenges such as identity preservation still exist, many remarkable achievements have been made.

So far, deep neural networks have been able to generate very photorealistic images, which are even difficult to distinguish by human. However, some limitations exist in the controllability of data generation and the diversity of generated data. One widely recognized disadvantage of neural networks is their "black box" nature, which means we don't know how to operate the intermediate output to modify some specific facial features. There have been some works try to infer the meanings of the latent vectors through the generated images [40]. But this additional operation has an upper limit of capability, which cannot disentangle facial features in the latent space completely. About image diversity, the facial appearance variation in the real world is unmeasurable, while the variability is limited with regard to synthetic data. If the face changes too much, it will be more difficult to preserve the identity.

8 Conclusion

In this paper, we give a literature review of face data augmentation and show the wide application and huge potential for various deep learning-based face detection and recognition tasks. We start with an introduction about the background and the concept of face data augmentation, which are followed by a brief review of related work. Then,

a detailed description of different transformations is presented to show the types supported for face data augmentation. Next, we present the commonly used methods of face data augmentation by introducing their principles and giving comparisons. The evaluation metrics for these methods are introduced subsequently. Finally, we discuss the challenges and opportunities. We hope this survey can give the beginners a general understanding of this field and give the related researchers the insight on future studies.

Compliances with ethical standard

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

A summary of recent works on face data augmentation is illustrated in the following table (Table 3). It summarizes these works from the transformation type, method and the evaluations they performed to test the capability of their algorithms for data augmentation. There is one notable thing that we only label the transformation types explicitly mentioned in original papers. Maybe the methods have the capability to be used for other transformations, but the authors did not mention in their original paper.

Table 3 Summary of recent works

Work	Alias	Supported transformation types							Method	Evaluated task
		HairStyle	Makeup	Accessory	Pose	Expression	Age	Others		
[167]	–					✓			GANs-based	Emotion classification
[34]	–				✓				2D model-based	–
[165]	–				✓				3D model-based	Face alignment
[97]	–				✓				3D model-based	Face recognition
[45]	–				✓				3D model-based	–
[11]	–				✓	✓		Illumination	GANs-based	–
[66]	DiscoGAN	✓		✓	✓			Gender	GANs-based	–
[19]	StarGAN	✓				✓	✓	Gender skin	GANs-based	–
[56]	pix2pix							Background	GANs-based	–
[87]	–	✓		✓		✓		Goatee	Generative-based	–
[88]	CoGAN	✓		✓		✓			GANs-based	–
[105]	Group-GAN FA-GAN F-GAN						✓		GANs-based	–
[156]	CAAE						✓		Generative-based	–
[4]	Age-cGAN						✓		GANs-based	–
[117]	–			✓		✓	✓	Beard gender	GANs-based	–
[67]	–		✓		✓	✓			GANs-based	–
[98]	–				✓	✓		Shape	3D model-based	Face recognition
[94]	–		✓	✓	✓			Illumination landmark perturbation	Image process 3D model-based	Face recognition
[147]	–					✓			Image process	–
[131]	–					✓			3D model-based	–

Table 3 (continued)

Work	Alias	HairStyle	Makeup	Accessory	Pose	Expression	Age	Others	Method	Evaluated task
[44]	–				√	√			3D model-based	–
[65]	–				√	√		Occlusion	3D model-based	3D face recognition
[162]	CDAAE					√			Generative-based	–
[150]	FVAE					√			Generative-based	–
[80]	DIAT			√		√	√	Gender	GANs-based	–
[136]	–				√	√		Illumination	Generative-based	–
[152]	–				√	√			GANs-based	Expression recognition
[46]	AttGAN	√		√		√	√	Beard eyebrows gender skin	GANs-based	–
[110]	GANimation					√			GANs-based	–
[127]	G2-GAN					√			GANs-based	–
[111]	GC-GAN					√			GANs-based	–
[29]	ExprGAN					√			GANs-based	Expression classification
[42]	–		√						Image process	–
[104]	–		√						Image process	–
[76]	–		√						Image process	–
[82]	BeautyGAN		√						GANs-based	–
[16]	–		√						GANs-based	–
[54]	TP-GAN				√				GANs-based	–
[151]	FF-GAN				√				GANs-based	–
[134]	DR-GAN				√				GANs-based	–
[142]	X2Face				√	√			Generative-based	–
[21]	–				√			Illumination	3D model-based	Face recognition
[73]	DC-IGN				√			Illumination	VAEs-based	–
[159, 160]	DA-GAN				√				3D Model and GANs	Face recognition
[63]	–	√					√	Beard	Image process	–
[18]	InfoGAN	√		√	√	√		Illumination shape	GANs-based	–
[141]	IPCGANs						√		GANs-based	Face recognition
[51]	CAPG-GAN				√				GANs-based	–
[27]	UV-GAN				√				3D model and GANs	Face recognition
[10]	CVAE-GAN				√	√			Generative-based	Face recognition
[36]	–				√	√		Illumination	3D model and GANs	Face recognition
[106]	CMMA			√				Beard shape	Generative-based	–
[148]	disCVAE	√		√		√	√	Gender	Generative-based	–

Table 3 (continued)

Work	Alias	Supported transformation types							Method	Evaluated task
		HairStyle	Makeup	Accessory	Pose	Expression	Age	Others		
[68]	Glow	√				√	√	Skin gender	Generative-based	–
[52]	IntroVAE				√			Gender	Generative-based	–
[157]	FSN						√		GANs-based	–
[163]	CMAAE-OR						√		GANs-based	–
[126]	Dual cGANs						√		GANs-based	–
[39]	D-GAN				√	√		Illumination	GANs-based	Expression classification
[72]	GAGAN				√	√			GANs-based	–
[14]	LB-GAN				√				GANs-based	–
[43]	–			√					Augmented reality	Face recognition
[144]	ReenactGAN					√			GANs-based	–
[90]	–						√		GANs-based	–
[109]	–					√			GANs-based	–
[115]	GANnotation				√	√			GANs-based	–
[81]	Wavelet GLCA-GAN						√		GANs-based	–
[132]	CR-GAN				√				GANs-based	–
[17]	TDB-GAN			√		√	√	Skin gender	GANs-based	–
[74]	Fader Networks			√		√	√	Gender	Generative-based	–
[146]	DNA-GAN	√		√		√		Illumination gender hat mustache	GANs-based	–
[161]	GeneGAN	√		√		√		Illumination	GANs-based	–

References

1. Agianpuye S, Minoi JL (2013) 3d facial expression synthesis: a survey. In: 2013 8th international conference on information technology in Asia (CITA). IEEE, pp 1–7

2. Alashkar T, Jiang S, Wang S, Fu Y (2017) Examples-rules guided deep neural network for makeup recommendation. In: Thirty-first AAAI conference on artificial intelligence

3. Alhaija HA, Mustikovela SK, Mescheder L, Geiger A, Rother C (2017) Augmented reality meets deep learning for car instance segmentation in urban scenes. In: British machine vision conference, vol 1, p 2

4. Antipov G, Baccouche M, Dugelay JL (2017) Face aging with conditional generative adversarial networks. In: 2017 IEEE international conference on image processing (ICIP). IEEE, pp 2089–2093

5. Antoniou A, Storkey A, Edwards H (2017) Data augmentation generative adversarial networks. ArXiv preprint arXiv:1711.04340

6. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: International conference on machine learning, pp 214–223

7. Azevedo P, Dos Santos TO, De Aguiar E (2016) An augmented reality virtual glasses try-on system. In: 2016 XVIII symposium on virtual and augmented reality (SVR). IEEE, pp 1–9

8. Banerjee S, Bernhard JS, Scheirer WJ, Bowyer KW, Flynn PJ (2017) Srefi: synthesis of realistic example face images. In: 2017 IEEE international joint conference on biometrics (IJCB). IEEE, pp 37–45

9. Banerjee S, Scheirer WJ, Bowyer KW, Flynn, PJ (2018) On hallucinating context and background pixels from a face mask using multi-scale gans. ArXiv preprint arXiv:1811.07104

10. Bao J, Chen D, Wen F, Li H, Hua G (2017) Cvae-gan: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE international conference on computer vision, pp 2745–2754

11. Bao J, Chen D, Wen F, Li H, Hua G (2018) Towards open-set identity preserving face synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6713–6722

12. Blanz V, Vetter T (2003) Face recognition based on fitting a 3d morphable model. IEEE Trans Pattern Anal Mach Intell 25(9):1063–1074

13. Blanz V, Vetter T et al (1999) A morphable model for the synthesis of 3d faces. Siggraph 99:187–194

14. Cao J, Hu Y, Yu B, He R, Sun Z (2018) Load balanced gans for multi-view face image synthesis. ArXiv preprint arXiv:1802.07447

15. Cao J, Hu Y, Zhang H, He R, Sun Z (2018) Learning a high fidelity pose invariant model for high-resolution face frontalization. In: Advances in neural information processing systems, pp 2872–2882

16. Chang H, Lu J, Yu F, Finkelstein A (2018) Pairedcyclegan: asymmetric style transfer for applying and removing makeup. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 40–48

17. Chen W, Xie X, Jia X, Shen L (2018) Texture deformation based generative adversarial networks for face editing. ArXiv preprint arXiv:1812.09832

18. Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I, Abbeel P (2016) Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Advances in neural information processing systems, pp 2172–2180

19. Choi Y, Choi M, Kim M, Ha JW, Kim S, Choo J (2018) Stargan: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8789–8797

20. Cootes TF, Edwards GJ, Taylor CJ (2001) Active appearance models. IEEE Trans Pattern Anal Mach Intell 6:681–685

21. Crispell D, Biris O, Crosswhite N, Byrne J, Mundy JL (2017) Dataset augmentation for pose and lighting invariant face recognition. ArXiv preprint arXiv:1704.04326

22. Cubuk E.D, Zoph B, Mane D, Vasudevan V, Le QV (2019) Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 113–123

23. Das D, Lee CSG (2018) Graph matching and pseudo-label guided deep unsupervised domain adaptation. In: International conference on artificial neural networks, pp 342–352

24. Das D, Lee CSG (2018) Sample-to-sample correspondence for unsupervised domain adaptation. Eng Appl Artif Intell 73:80–91

25. Das D, Lee CSG (2018) Unsupervised domain adaptation using regularized hyper-graph matching. In: Computer vision and pattern recognition

26. Das D, Lee CSG (2019) Zero-shot image recognition using relational matching, adaptation and calibration. In: Computer vision and pattern recognition

27. Deng J, Cheng S, Xue N, Zhou Y, Zafeiriou S (2018) Uv-gan: Adversarial facial uv map completion for pose-invariant face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7093–7102

28. Di X, Sindagi VA, Patel VM (2018) Gp-gan: Gender preserving gan for synthesizing faces from landmarks. In: 2018 24th international conference on pattern recognition (ICPR). IEEE, pp 1079–1084

29. Ding H, Sricharan K, Chellappa R (2018) Exprgan: facial expression editing with controllable expression intensity. In: Thirty-second AAAI conference on artificial intelligence

30. Dinh L, Krueger D, Bengio Y (2014) Nice: non-linear independent components estimation. ArXiv preprint arXiv:1410.8516

31. Dosovitskiy A, Fischer P, Ilg E, Hausser P, Hazirbas C, Golkov V, Van Der Smagt P, Cremers D, Brox T (2015) Flownet: learning optical flow with convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 2758–2766

32. Faceapp: Transform your face. https://www.faceapp.com/ (2018)

33. Feng ZH, Hu G, Kittler J, Christmas W, Wu XJ (2015) Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting. IEEE Trans Image Process 24(11):3425–3440

34. Feng ZH, Kittler J, Christmas W, Huber P, Wu XJ (2017) Dynamic attention-controlled cascaded shape regression exploiting training data augmentation and fuzzy-set sample weighting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2481–2490

35. Flynn M (2016) Generating faces with deconvolution networks. https://zo7.github.io/blog/2016/09/25/generating-faces.html

36. Gecer B, Bhattarai B, Kittler J, Kim TK (2018) Semi-supervised adversarial learning to generate photorealistic face images of new identities from 3d morphable model. In: Proceedings of the European conference on computer vision (ECCV), pp 217–234

37. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680

38. Grover A, Dhar M, Ermon S (2018) Flow-gan: Combining maximum likelihood and adversarial learning in generative models. In: Thirty-second AAAI conference on artificial intelligence

39. Gu G, Kim ST, Kim K, Baddar WJ, Ro YM (2017) Differential generative adversarial networks: synthesizing non-linear facial variations with limited number of training data. ArXiv preprint arXiv:1711.10267

40. Guan S (2018) Tl-gan: transparent latent-space gan. https://github.com/SummitKwan/transparent_latent_gan

41. Gulrajani I, Kumar K, Ahmed F, Taiga AA, Visin F, Vazquez D, Courville A (2016) Pixelvae: a latent variable model for natural images. ArXiv preprint arXiv:1611.05013

42. Guo D, Sim T (2009) Digital face makeup by example. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE, pp 73–79

43. Guo J, Zhu X, Lei Z, Li SZ (2018) Face synthesis for eyeglass-robust face recognition. In: Chinese conference on biometric recognition. Springer, pp 275–284

44. Guo Y, Cai J, Jiang B, Zheng J et al (2018) Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images. IEEE Trans Pattern Anal Mach Intell 41(6):1294–1307

45. Hassner T, Harel S, Paz E, Enbar R (2015) Effective face frontalization in unconstrained images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4295–4304

46. He Z, Zuo W, Kan M, Shan S, Chen X (2019) Attgan: facial attribute editing by only changing what you want. IEEE Trans Image Process 28(11):5464–5478

47. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in neural information processing systems, pp 6626–6637

48. Hong S, Im W, Ryu J, Yang HS (2017) Sspp-dan: Deep domain adaptation network for face recognition with single sample per person. In: 2017 IEEE international conference on image processing (ICIP). IEEE, pp 825–829

49. Hu G, Peng X, Yang Y, Hospedales TM, Verbeek J (2018) Frankenstein: learning deep face representations using small data. IEEE Trans Image Process 27(1):293–303

50. Hu G, Yan F, Chan C.H, Deng W, Christmas W, Kittler J, Robertson NM (2016) Face recognition using a unified 3d morphable model. In: European conference on computer vision. Springer, pp 73–89

51. Hu Y, Wu X, Yu B, He R, Sun Z (2018) Pose-guided photo-realistic face rotation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8398–8406

52. Huang H, He R, Sun Z, Tan T et al (2018) Introvae: introspective variational autoencoders for photographic image synthesis. In: Advances in neural information processing systems, pp 52–63

53. Huang H, Yu PS, Wang C (2018) An introduction to image synthesis with generative adversarial nets. ArXiv preprint arXiv:1803.04469

54. Huang R, Zhang S, Li T, He R (2017) Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis. In: Proceedings of the IEEE international conference on computer vision, pp 2439–2448

55. Huber P, Hu G, Tena R, Mortazavian P, Koppen P, Christmas WJ, Ratsch M, Kittler J (2016) A multiresolution 3d morphable face model and fitting framework. In: Proceedings of the 11th international joint conference on computer vision, imaging and computer graphics theory and applications

56. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134

57. Javornik A, Rogers Y, Moutinho AM, Freeman R (2016) Revealing the shopper experience of using a" magic mirror" augmented reality make-up application. In: Conference on designing interactive systems, vol 2016. Association for Computing Machinery (ACM), pp 871–882

58. Johnson J, Alahi A, Fei-Fei L (2016) Perceptual losses for real-time style transfer and super-resolution. In: European conference on computer vision. Springer, pp 694–711

59. Juefei-Xu F, Dey R, Bodetti V, Savvides M (2018) Rankgan: a maximum margin ranking gan for generating faces. In: Proceedings of the Asian conference on computer vision (ACCV), vol 4

60. Jung A (2017) imgaug. https://github.com/aleju/imgaug

61. Karras T, Aila T, Laine S, Lehtinen J (2018) Progressive growing of gans for improved quality, stability, and variation. In: ICLR

62. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4401–4410

63. Kemelmacher-Shlizerman I (2016) Transfiguring portraits. ACM Trans Graph (TOG) 35(4):94

64. Kemelmacher-Shlizerman I, Suwajanakorn S, Seitz SM (2014) Illumination-aware age progression. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3334–3341

65. Kim D, Hernandez M, Choi J, Medioni G (2017) Deep 3d face identification. In: 2017 IEEE international joint conference on biometrics (IJCB). IEEE, pp 133–142

66. Kim T, Cha M, Kim H, Lee JK, Kim J (2017) Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th international conference on machine learning, volume 70, pp 1857–1865. JMLR.org

67. Kim T, Kim B, Cha M, Kim J (2017) Unsupervised visual attribute transfer with reconfigurable generative adversarial networks. ArXiv preprint arXiv:1707.09798

68. Kingma DP, Dhariwal P (2018) Glow: Generative flow with invertible 1x1 convolutions. In: Advances in neural information processing systems, pp 10236–10245

69. Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: Proceedings of the 2nd international conference on learning representations (ICLR)

70. Kitanovski V, Izquierdo E (2011) Augmented reality mirror for virtual facial alterations. In: 2011 18th IEEE international conference on image processing. IEEE, pp 1093–1096

71. Kortylewski A, Schneider A, Gerig T, Egger B, Morel-Forster A, Vetter T (2018) Training deep face recognition systems with synthetic data. ArXiv preprint arXiv:1802.05891

72. Kossaifi J, Tran L, Panagakis Y, Pantic M (2018) Gagan: geometry-aware generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 878–887

73. Kulkarni TD, Whitney WF, Kohli P, Tenenbaum J (2015) Deep convolutional inverse graphics network. In: Advances in neural information processing systems, pp 2539–2547

74. Lample G, Zeghidour N, Usunier N, Bordes A, Denoyer L et al (2017) Fader networks: manipulating images by sliding attributes. In: Advances in neural information processing systems, pp 5967–5976

75. Larsen ABL, Sønderby SK, Larochelle H, Winther O (2015) Autoencoding beyond pixels using a learned similarity metric. In: 33rd International conference on machine learning

76. Lee JY, Kang HB (2016) A new digital face makeup method. In: 2016 IEEE international conference on consumer electronics (ICCE). IEEE, pp 129–130

77. Lemley J, Bazrafkan S, Corcoran P (2017) Smart augmentation learning an optimal data augmentation strategy. IEEE Access 5:5858–5869

78. Leng B, Yu K, Jingyan Q (2017) Data augmentation for unbalanced face recognition training sets. Neurocomputing 235:10–14

79. Li L, Peng Y, Qiu G, Sun Z, Liu S (2018) A survey of virtual sample generation technology for face recognition. Artif Intell Rev 50(1):1–20

80. Li M, Zuo W, Zhang D (2016) Deep identity-aware transfer of facial attributes. ArXiv preprint arXiv:1610.05586

81. Li P, Hu Y, He R, Sun Z (2019) Global and local consistent wavelet-domain age synthesis. IEEE Trans Inf Forensics Secur 14(11):2943–2957

82. Li T, Qian R, Dong C, Liu S, Yan Q, Zhu W, Lin L (2018) Beautygan: instance-level facial makeup transfer with deep generative adversarial network. In: 2018 ACM multimedia conference on multimedia conference. ACM, pp 645–653

83. Liu B, Wang X, Dixit M, Kwitt R, Vasconcelos N (2018) Feature space transfer for data augmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9090–9098

84. Liu F, Zhu R, Zeng D, Zhao Q, Liu X (2018) Disentangling features in 3d face shapes for joint face reconstruction and recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5216–5225

85. Liu L, Xing J, Liu S, Xu H, Zhou X, Yan S (2014) Wow! you are so beautiful today!. ACM Trans Multimed Comput Commun Appl (TOMM) 11(1s):20

86. Liu L, Zhang H, Ji Y, Wu QJ (2019) Toward ai fashion design: an attribute-gan model for clothing match. Neurocomputing 341:156–167

87. Liu MY, Breuel T, Kautz J (2017) Unsupervised image-to-image translation networks. In: Advances in neural information processing systems, pp 700–708

88. Liu MY, Tuzel O (2016) Coupled generative adversarial networks. In: Advances in neural information processing systems, pp 469–477

89. Liu S, Ou X, Qian R, Wang W, Cao X (2016) Makeup like a superstar: deep localized makeup transfer network. In: Proceedings of the twenty-fifth international joint conference on artificial intelligence, IJCAI'16. AAAI Press, pp 2568–2575

90. Liu Y, Li Q, Sun Z (2019) Attribute enhanced face aging with wavelet-based generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition

91. Liu Z, Luo P, Wang X, Tang X (2015) Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision, pp 3730–3738
92. Lu Y, Tai YW, Tang CK (2018) Attribute-guided face generation using conditional cyclegan. In: Proceedings of the European conference on computer vision (ECCV), pp 282–297
93. Lv JJ, Cheng C, Tian GD, Zhou XD, Zhou X (2016) Landmark perturbation-based data augmentation for unconstrained face recognition. Signal Process Image Commun 47:465–475
94. Lv JJ, Shao XH, Huang JS, Zhou XD, Zhou X (2017) Data augmentation for face recognition. Neurocomputing 230:184–196
95. Makhzani A, Shlens J, Jaitly N, Goodfellow I, Frey B (2015) Adversarial autoencoders. ArXiv preprint arXiv:1511.05644
96. Mash R, Borghetti B, Pecarina J (2016) Improved aircraft recognition for aerial refueling through data augmentation in convolutional neural networks. In: International symposium on visual computing. Springer, pp 113–122
97. Masi I, Rawls S, Medioni G, Natarajan P (2016) Pose-aware face recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4838–4846
98. Masi I, Trãn AT, Hassner T, Leksut JT, Medioni G (2016) Do we really need to collect millions of faces for effective face recognition? In: European conference on computer vision. Springer, pp 579–596
99. Matthews I, Xiao J, Baker S (2007) 2d vs. 3d deformable face models: representational power, construction, and real-time fitting. Int J Comput Vision 75(1):93–113
100. Menze M, Geiger, A (2015) Object scene flow for autonomous vehicles. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3061–3070
101. Mirza M, Osindero, S (2014) Conditional generative adversarial nets. ArXiv preprint arXiv:1411.1784
102. Moniz JRA, Beckham C, Rajotte S, Honari S, Pal C (2018) Unsupervised depth estimation, 3d face rotation and replacement. In: Advances in neural information processing systems, pp 9759–9769
103. Nguyen TV, Liu L (2017) Smart mirror: Intelligent makeup recommendation and synthesis. In: Proceedings of the 25th ACM international conference on multimedia. ACM, pp 1253–1254
104. Oo WY (2016) Digital makeup face generation. https://web.stanford.edu/class/ee368/Project_Autumn_1516/Reports/Oo.pdf
105. Palsson S, Agustsson E, Timofte R, Van Gool L (2018) Generative adversarial style transfer networks for face aging. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 2084–2092
106. Pandey G, Dukkipati A (2017) Variational methods for conditional multimodal deep learning. In: 2017 international joint conference on neural networks (IJCNN). IEEE, pp 308–315
107. Parkhi OM, Vedaldi A, Zisserman A et al (2015) Deep face recognition. In: The British machine vision conference (BMVC), vol 1, p 6
108. Perarnau G, Van De Weijer J, Raducanu B, Álvarez JM (2016) Invertible conditional gans for image editing. In: NIPS 2016 workshop on adversarial training
109. Pham HX, Wang Y, Pavlovic V (2018) Generative adversarial talking head: Bringing portraits to life with a weakly supervised neural network. ArXiv preprint arXiv:1803.07716
110. Pumarola A, Agudo A, Martinez AM, Sanfeliu A, Moreno-Noguer F (2018) Ganimation: anatomically-aware facial animation from a single image. In: Proceedings of the European conference on computer vision (ECCV), pp 818–833

111. Qiao F, Yao N, Jiao Z, Li Z, Chen H, Wang H (2018) Geometry-contrastive gan for facial expression transfer. ArXiv preprint arXiv:1802.01822
112. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. ArXiv preprint arXiv:1511.06434
113. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. In: Advances in neural information processing systems, pp 2234–2242
114. Salimans T, Karpathy A, Chen X, Kingma DP (2017) Pixelcnn++: a pixelcnn implementation with discretized logistic mixture likelihood and other modifications. In: ICLR
115. Sanchez E, Valstar M (2018) Triple consistency loss for pairing distributions in gan-based face synthesis. ArXiv preprint arXiv:1811.03492
116. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 815–823
117. Shen W, Liu R (2017) Learning residual images for face attribute manipulation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4030–4038
118. Shen Y, Luo P, Yan J, Wang X, Tang X (2018) Faceid-gan: learning a symmetry three-player gan for identity-preserving face synthesis. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 821–830
119. Shen Y, Zhou B, Luo P, Tang X (2018) Facefeat-gan: a two-stage approach for identity-preserving face synthesis. ArXiv preprint arXiv:1812.01288
120. Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R (2017) Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2107–2116
121. Shu X, Tang J, Lai H, Liu L, Yan S (2015) Personalized age progression with aging dictionary. In: Proceedings of the IEEE international conference on computer vision, pp 3970–3978
122. Shu Z, Sahasrabudhe M, Alp Guler R, Samaras D, Paragios N, Kokkinos I (2018) Deforming autoencoders: unsupervised disentangling of shape and appearance. In: Proceedings of the European conference on computer vision (ECCV), pp 650–665
123. Shu Z, Yumer E, Hadap S, Sunkavalli K, Shechtman E, Samaras D (2017) Neural face editing with intrinsic image disentangling. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5541–5550
124. Sixt L, Wild B, Landgraf T (2018) Rendergan: generating realistic labeled data. Front Robot AI 5:66
125. Sohn K, Lee H, Yan X (2015) Learning structured output representation using deep conditional generative models. In: Advances in neural information processing systems, pp 3483–3491
126. Song J, Zhang J, Gao L, Liu X, Shen HT (2018) Dual conditional gans for face aging and rejuvenation. In: IJCAI, pp 899–905
127. Song L, Lu Z, He R, Sun Z, Tan T (2018) Geometry guided adversarial facial expression synthesis. In: 2018 ACM multimedia conference on multimedia conference. ACM, pp 627–635
128. Suo J, Zhu SC, Shan S, Chen X (2010) A compositional and dynamic model for face aging. IEEE Trans Pattern Anal Mach Intell 32(3):385–401
129. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1701–1708

130. Taylor L, Nitschke G (2018) Improving deep learning using generic data augmentation. In: 2018 IEEE symposium series on computational intelligence (SSCI). IEEE

131. Thies J, Zollhöfer M, Nießner M, Valgaerts L, Stamminger M, Theobalt C (2015) Real-time expression transfer for facial reenactment. ACM Trans Graph 34(6):183-1

132. Tian Y, Peng X, Zhao L, Zhang S, Metaxas DN (2018) Cr-gan: learning complete representations for multi-view generation. In: International joint conference on artificial intelligence (IJCAI)

133. Tran L, Liu X (2018) Nonlinear 3d face morphable model. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7346–7355

134. Tran L, Yin X, Liu X (2017) Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1415–1424

135. van Krevelen DWF, Poelman R (2010) A survey of augmented reality technologies, applications and limitations. Int J Virtual Real 9(2):1–20

136. Van den Oord A, Kalchbrenner N, Espeholt L, Vinyals O, Graves A et al (2016) Conditional image generation with pixelcnn decoders. In: Advances in neural information processing systems, pp 4790–4798

137. Van Den Oord A, Kalchbrenner N, Kavukcuoglu K (2016) Pixel recurrent neural networks. In: Proceedings of the 33rd international conference on international conference on machine learning-volume 48, ICML'16. JMLR.org, pp 1747–1756

138. Volpi R, Namkoong H, Sener O, Duchi JC, Murino V, Savarese S (2018) Generalizing to unseen domains via adversarial data augmentation. In: Advances in neural information processing systems, pp 5334–5344

139. Wang J, Perez L (2017) The effectiveness of data augmentation in image classification using deep learning. Convolutional Neural Netw Vis Recognit 11

140. Wang W, Cui Z, Yan Y, Feng J, Yan S, Shu X, Sebe N (2016) Recurrent face aging. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2378–2386

141. Wang Z, Tang X, Luo W, Gao S (2018) Face aging with identity-preserved conditional generative adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7939–7947

142. Wiles O, Sophia Koepke A, Zisserman A (2018) X2face: A network for controlling face generation using images, audio, and pose codes. In: Proceedings of the European conference on computer vision (ECCV), pp 670–686

143. Winston H (2018) Investigating data augmentation strategies for advancing deep learning training. https://winstonhsu.info/wp-content/uploads/2018/03/gtc18-data_aug-180326.pdf

144. Wu W, Zhang Y, Li C, Qian C, Change Loy C (2018) Reenactgan: learning to reenact faces via boundary transfer. In: Proceedings of the European conference on computer vision (ECCV), pp 603–619

145. Wu X, He R, Sun Z, Tan T (2018) A light cnn for deep face representation with noisy labels. IEEE Trans Inf Forensics Secur 13(11):2884–2896

146. Xiao T, Hong J, Ma J (2018) Dna-gan: learning disentangled representations from multi-attribute images. In: International conference on learning representations workshop 2018

147. Xie W, Shen L, Yang M, Jiang J (2018) Facial expression synthesis with direction field preservation based mesh deformation and lighting fitting based wrinkle mapping. Multimed Tools Appl 77(6):7565–7593

148. Yan X, Yang J, Sohn K, Lee H (2016) Attribute2image: conditional image generation from visual attributes. In: European conference on computer vision. Springer, pp 776–791

149. Yang H, Huang D, Wang Y, Jain AK (2018) Learning face age progression: a pyramid architecture of gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 31–39

150. Yeh R, Liu Z, Goldman DB, Agarwala A (2016) Semantic facial expression editing using autoencoded flow. ArXiv preprint arXiv:1611.09961

151. Yin X, Yu X, Sohn K, Liu X, Chandraker M (2017) Towards large-pose face frontalization in the wild. In: Proceedings of the IEEE international conference on computer vision, pp 3990–3999

152. Zhang F, Zhang T, Mao Q, Xu C (2018) Joint pose and expression modeling for facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3359–3368

153. Zhang H, Sun Y, Liu L, Wang X, Li L, Liu W (2018) Clothingout: a category-supervised GAN model for clothing segmentation and retrieval. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3691-y

154. Zhang H, Xu T, Li H, Zhang S, Wang X, Huang X, Metaxas DN (2017) Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 5907–5915

155. Zhang L, Samaras D (2006) Face recognition from a single training image under arbitrary unknown lighting using spherical harmonics. IEEE Trans Pattern Anal Mach Intell 28(3):351–363

156. Zhang Z, Song Y, Qi H (2017) Age progression/regression by conditional adversarial autoencoder. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5810–5818

157. Zhao J, Cheng Y, Cheng Y, Yang Y, Zhao F, Li J, Liu H, Yan S, Feng J (2019) Look across elapse: disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. Proc AAAI Conf Artif Intell 33:9251–9258

158. Zhao J, Cheng Y, Xu Y, Xiong L, Li J, Zhao F, Jayashree K, Pranata S, Shen S, Xing J et al (2018) Towards pose invariant face recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2207–2216

159. Zhao J, Xiong L, Jayashree PK, Li J, Zhao F, Wang Z, Pranata PS, Shen PS, Yan S, Feng J (2017) Dual-agent gans for photorealistic and identity preserving profile face synthesis. In: Advances in neural information processing systems, pp 66–76

160. Zhao J, Xiong L, Li J, Xing J, Yan S, Feng J (2018) 3d-aided dual-agent gans for unconstrained face recognition. IEEE Trans Pattern Anal Mach Intell 41(10):2380–2394

161. Zhou S, Xiao T, Yang Y, Feng D, He Q, He W (2017) Genegan: learning object transfiguration and attribute subspace from unpaired data. In: Proceedings of the British machine vision conference 2017

162. Zhou Y, Shi BE (2017) Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In: 2017 seventh international conference on affective computing and intelligent interaction (ACII). IEEE, pp 370–376

163. Zhu H, Zhou Q, Zhang J, Wang JZ (2018) Facial aging and rejuvenation by conditional multi-adversarial autoencoder with ordinal regression. ArXiv preprint arXiv:1804.02740

164. Zhu J.Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

165. Zhu X, Lei Z, Liu X, Shi H, Li SZ (2016) Face alignment across large poses: a 3d solution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 146–155

166. Zhu X, Lei Z, Yan J, Yi D, Li SZ (2015) High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 787–796

167. Zhu X, Liu Y, Li J, Wan T, Qin Z (2018) Emotion classification with data augmentation using generative adversarial networks. In: Pacific-Asia conference on knowledge discovery and data mining. Springer, pp 349–360

168. Zhuang L, Yang AY, Zhou Z, Shankar Sastry S, Ma Y (2013) Single-sample face recognition with image corruption and misalignment via sparse illumination transfer. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3546–3553