Abstract—Facial Attribute Manipulation (FAM) aims to aesthetically modify a given face image to render desired attributes, which has received significant attention due to its broad practical applications ranging from digital entertainment to biometric forensics. In the last decade, with the remarkable success of Generative Adversarial Networks (GANs) in synthesizing realistic images, numerous GAN-based models have been proposed to solve FAM with various problem formulation approaches and guiding information representations. This paper presents a comprehensive survey of GAN-based FAM methods with a focus on summarizing their principal motivations and technical details. The main contents of this survey include: (i) an introduction to the research background and basic concepts related to FAM, (ii) a systematic review of GAN-based FAM methods in three main categories, and (iii) an in-depth discussion of important properties of FAM methods, open issues, and future research directions. This survey not only builds a good starting point for researchers new to this field but also serves as a reference for the vision community.

Index Terms—Generative adversarial networks, image translation, facial attribute manipulation.

I. INTRODUCTION

Facial attributes describe the semantic properties of human faces in an interpretable manner. As soft biometrics presenting the classifiable visual characteristic of faces with rich individual variations, facial attributes have been playing a vital role in research on many computer vision tasks, such as face recognition [1], [2], [3] and face retrieval [4], [5], [6]. With the development of deep generative models, facial attribute manipulation (FAM) has attracted growing research attention due to its profound theoretical significance and broad practical applications.

Existing FAM methods aim to modify input face images to render target attributes while ensuring the visual quality of manipulation results (see Fig. 1). Early research on FAM can be traced back to the 1990s [7], [8], where the modeling of faces largely involves complex prior knowledge specific to the target task. As a class of parameterized models describing the principal variations of face shape, 3D Morphable Models (3DMMs) [9] are widely used in many studies to manipulate attributes related to facial geometry (e.g., pose [10], [11], [12] and expression [13], [14]). In addition, the Facial Action Coding System (FACS) [15], a framework that decomposes and characterizes visually perceptible facial muscle movements, has been widely used in studies related to facial expression analysis [16], [17] and manipulation [18], [19]. Furthermore, multi-linear tensor models [20], [21] could also be used for performing FAM, as they can separate and parameterize the space of facial semantic variations due to different attributes. Although FAM can be performed by these model-driven approaches, the algorithm pipelines are usually complicated, and synthesized results are of low visual quality (see Fig. 1(a)). To solve this problem, subsequent studies [22], [23], [24] adopt data-driven deep models for feature extraction and image translation, which can generate high-quality images in an end-to-end manner.

In the last decade, the development of Generative Adversarial Networks (GANs) [27] have made significant advances in improving the visual quality of synthesized images [28], [29], [30], [31], [32], and GAN-based FAM has received much attention from the community of computer vision. Due to the intrinsic
complexity of face images and semantic variations of facial attributes, FAM is one of the most important applications of GANs for studying controllable and diverse image translation. It not only serves as a common benchmark for general-purpose image translation models based on GANs [33], [34], [35], but also motivates numerous solutions based on various problem settings [36], [37], [38], [39], [40], [41]. Moreover, FAM also has great practical significance as it is closely related to many real-world applications, such as digital entertainment (editing photos posted on social media), face recognition (normalizing pose and expression), and information forensics (augmenting training data for DeepFake detection).

In this paper, we present a comprehensive survey of existing GAN-based FAM methods with a focus on discussing the motivations and features, as well as summarizing the similarities and differences to give a whole picture of representative approaches in this field. Moreover, we also identify challenges and open issues of existing studies and forecast future research directions. Although some surveys [42], [43] have also discussed GAN-based FAM methods, they do not include those built on pre-trained generative priors, which have been extensively studied in recent years. On the other hand, the review [44] focuses on GAN inversion methods, and FAM is only considered as one of the downstream applications. Moreover, FAM is also discussed in [45] and [46], but the main themes are on data augmentation and DeepFake detection, respectively.

The rest of this article is organized as follows. In Section II, we define important concepts in FAM and the scope of this survey, and then categorize existing GAN-based FAM methods in terms of methodology. To conduct a self-contained review of the literature, the basic formulation of GAN models, the commonly-used datasets, and metrics for quantitative evaluation are introduced in Section III. Representative GAN-based FAM methods are then systematically reviewed in Sections IV, V, and VI. In Section VII, we summarize the important properties of FAM and how they are approached by GAN-based methods. We discuss topics closely related to FAM as well as challenges and future research directions in Section VIII.

II. PRELIMINARIES

A. Facial Attribute Manipulation

In recent years, numerous GAN models have been developed for image-to-image translation (I2I) [43], which is the broadest concept among all topics related to image manipulation. On the other hand, face image translation aims to change the holistic appearance, including identity swapping [47], [48] and make-up transfer [49], [50]. Facial attribute manipulation, the research topic of this survey, further constrains the manipulated semantics to facial attributes.

In this paper, we define facial attributes as the inherent properties of human faces which are categorical and interpretable. They describe generic visual characteristics of facial components (including hair and accessories) or soft-biometrics (e.g., age, gender, and race), which not only can be categorized into discrete classes but also contain large individual variations in terms of textural details. As such, FAM can be clearly distinguished from the following face image translation tasks:

1) Identity information editing: In this paper, identity information is not considered as a facial attribute since it does not present an interpretable or categorical feature of human faces. Therefore, related tasks such as identity anonymization [51], [52] and face swapping [47], [48] are beyond the scope of FAM.

2) Image quality enhancement: Although the prior knowledge of facial biometrics is widely incorporated when improving the quality of portrait images, such as in image super-resolution [53], [54] or image restoration [55], images are translated in terms of the quality of textural details or overall perception, which are not considered as facial attributes.

3) Face image stylization: Similar to image quality, artistic styles are not biological properties specific to face images, and thus can not be treated as facial attributes. Therefore, face image cartoonization [56], [57] and colorization [58], [59] are out of the scope of this survey.

Fig. 2 presents sample results and illustrates the difference between FAM and other face image translation tasks.

B. Taxonomy

A typical GAN model consists of two networks, a generator $G$ mapping an input variable $z \sim p_z$ to the output image ($p_z$ denotes the prior distribution), and a discriminator $D$ distinguishing generated images $G(z)$ from real ones $x$. These two modules are adversarially trained to encourage $G$ to produce visually plausible output images. For a GAN-based FAM method where the input is a face image $x$, an encoding process $E$ is usually required to project $x$ into a latent space $Z$ and obtain its embedded representation $z = E(x) \in Z$ for editing. Based on how $z$ is obtained
and manipulated to render the desired attribute change, we broadly categorize GAN-based FAM methods into three main groups: image domain translation-based methods, facial semantic decomposition-based methods, and latent space navigation-based approaches (Fig. 3).

Image domain translation-based methods consider FAM as an image-to-image translation (I2I) problem, where face images are grouped into distinctive categories (i.e., image domains) according to their attributes, and domain-level mapping functions are learned to perform image translation. Early work in this category mainly focuses on manipulating only one facial attribute with a single model [62], [63], [64], and subsequent schemes propose to scale to multi-attribute editing by incorporating various kinds of conditional information (denoted as e in Fig. 3(a)) [33], [34], [60], [65], [66].

Numerous FAM methods based on semantic segmentation [36], [37], [38], [39], [40], [41], [61], [67], [68], [69], [70], [71], [72], [73], [74], [75] decompose an input image \( x \) into separate latent spaces (instead of computing a joint latent representation \( z \)), where each embedding in the latent spaces is responsible for controlling different image semantics of \( x \). Compared to domain-level image translation, this class of methods successfully captures the intra-domain image variation, i.e., the difference in detailed style information of an attribute (e.g., the appearance of bangs in Fig. 3(b)), which enables generating diverse FAM results with fine-grained controllability and high flexibility.

A number of methods based on latent space navigation [35], [76], [77], [78], [79] achieve FAM by traversing an embedding \( z = G_{inv} (x) \) \((G_{inv} \) denotes the GAN inversion operation [44]) within the latent space \( Z \) of a pre-trained generator \( G^* \) such that the decoding result of the translated latent code \( z' \), i.e., \( G^*(z') \), can present the desired changes of facial attributes. This category of methods has gained significant research interest with the emergence of large-scale GAN models (e.g., BigGAN [80], PGGAN [29], and StyleGANs [30], [31], [32]), as they facilitate efficient manipulation of high-resolution face images (up to \( 1024 \times 1024 \)) by utilizing pre-trained unconditional generators without the need for re-training or modifying the deep network.

III. FORMULATION, DATASET, AND METRICS

For a self-contained literature review on FAM, we introduce basic formulations of GANs, datasets in GAN-based FAM methods, and metrics for quantitative evaluation.

A. Basic Formulations of GANs

GANs are a class of deep neural networks that learn to estimate the mapping \( G : Z \rightarrow X \) with a generator network \( G \), where \( Z \) is a latent space with a tractable prior \( p_z \) and \( X \) is the target space with an intractable data distribution \( p_x \). Moreover, a discriminator \( D \) is adopted to distinguish \( p_y \) from \( p_z \), where \( p_y \) denotes the distribution of the output of \( G \). The parameters of \( G \) and \( D \) are optimized alternatively via an adversarial process formulated by

\[
\min_G \max_D L_{GAN} = \mathbb{E}_x [\log D (x)] + \mathbb{E}_{z \sim p_z} [\log (1 - D (G (z)))]
\]

(1)

It is well-known that care must be exercised when training a GAN model to deal with the vanishing gradients [27] and mode collapse [81] issues. Therefore, many subsequent methods have been proposed to stabilize the training process and improve the quality and diversity of generated images by developing advanced objective functions [81], [82], [83] and network structures [28], [29]. As shown in Fig. 4(a), the generator of these GAN models directly takes the latent code \( z \) as input to the stacked convolutional layers and produces the output image, which we term as ‘traditional generators’.

In contrast, the ‘style-based generators’ [30], [31], [32] take a learnable constant tensor \( \mathbf{c} \) as the input to the synthesis network \( g \) (see Fig. 4(b)). To control the semantics of generation results, a latent code \( z \in Z \) is first projected into an intermediate latent space \( W \) via a mapping network \( f \), i.e., \( w = f (z) \in W \). Next, \( w \) is replicated by \( N \) times and fed into separate affine transformation networks \( A_i \) \((i = 1, 2, \ldots, N) \) to compute coefficients \( s_i = A_i (w) \) (i.e., style codes), where \( N \) is dependent on the output resolution (two times the number of up-samples blocks) in StyleGAN1 [30] and StyleGAN2 [31]. On the other hand, this
parameter $N$ has no direct effect on the size of generated images in StyleGAN3 [32]. These style codes are then used to modulate feature maps at different levels in $g$. Such a modulation mechanism is implemented via adaptive instance normalization [84] in StyleGAN1 [30], and replaced by style modulation layers in StyleGAN2 [31] to remove artifacts.

While the $W$ space contains more disentangled features than $Z$ due to the mapping function $f$ [30], [79], the representation model still cannot accurately reconstruct a wide range of images through GAN inversion. Therefore, the $W^+$ space, where different latent codes $w$ are fed into each of the affine transformation blocks, is proposed [85], [86] to alleviate distortion and improve the faithfulness of reconstruction. Furthermore, the $S$ space (also referred to as the ‘style space’) is spanned by channel-wise style coefficients obtained by forwarding $w$ through different learned affine transformations, which is proven to be able to achieve better spatial disentanglement in image editing [87], [88].

It is also worth mentioning that numerous extensions of StyleGAN have been proposed recently to improve image quality under limited training data (StyleGAN-ADA [89]) or specific synthesis tasks (StyleGAN-V [90], StyleGAN-T [91], and StyleGAN-NADA [92]).

B. Facial Attribute Datasets

Numerous datasets have been used for FAM including:

- **FaceTracer** [93] contains 15,000 face images collected in uncontrolled environments. A subset of these images is manually labeled with 7 groups of facial attributes, consisting of 19 labels in total.

- **PubFig** [94] contains 58,797 face images of 200 identities collected from the Internet. Each image in PubFig is labeled with 73 facial attributes.

- **LFWA** [95] is created by annotating 40 facial attributes with barycentric labels of images in LFW [96]. It consists of 13,233 facial images of 5,729 unique identities collected from the Internet.

- **CelebA** [95] is the most widely used large-scale face dataset in GAN-based FAM studies due to its diversity of attribute labels and large variation of image content. It consists of 202,599 celebrity face images from CelebFaces [97] with binary labels for 40 facial attributes and 10,177 identities. **CelebA-HQ** [29] is a high-resolution subset (1024 × 1024) of CelebA that contains 30,000 images. Additionally, **CelebAMask-HQ** [40] provides pixel-level annotations for 19 classes of facial components and accessories in CelebA-HQ images that have been down-sampled to 512 × 512 resolution.

- **FFHQ** [30] is a high-quality face dataset that consists of 70,000 images with a resolution of 1024 × 1024. The images were initially collected from Flickr, then manually reviewed to remove low-quality samples, and normalized using dlib [98]. Additionally, a number of datasets have also been introduced specifically for manipulating individual facial attributes of FAM models. For example, **FG-Net** [99], **MORPH** [100], **CACD** [101], **UTKFace** [102], and **FFHQ-Aging** [103] are proposed to train FAM methods for age progression and regression. For facial expression manipulation, **BU-3DFE** [104], **Bosphorus** [105], **RaFD** [106], **FaceWareHouse** [107], and **EmotioNet** [108] are created; There are also datasets proposed for transferring the style of facial makeups, such as **MakeUp-Transfer** [109], **Makeup-Wild** [50], **LADN** [49], and **Beauty-Face** [110].

C. Evaluation Metrics

We summarize evaluation metrics widely used for FAM in four aspects: realism of manipulated images, accuracy of attribute translation, consistency of image semantics, and properties of latent spaces.

1) **Realism of Manipulated Images**: The IS, FID, KID, and LPIPS metrics are commonly adopted for evaluating the realism of the output of FAM methods. Other metrics such as Conditional IS [68] and sliced Wasserstein discrepancy (SWD) [111] have also been utilized for measuring the visual quality.

**Inception Score (IS)** [112] serves as an alternative to human annotators for measuring both the diversity and interpretability of images generated by GANs. It calculates the KL-Divergence between the marginal and conditional label distributions, which are computed by a pre-trained Inception-v3 network [113].

**Fréchet Inception Distance (FID)** [114] assesses the discrepancy between $p_z$ and $p_x$ in terms of the Fréchet Distance, which is calculated based on the features extracted by a pre-trained Inception-v3 network.

**Kernel Inception Distance (KID)** [115] evaluates the maximum mean discrepancy (MMD) between the Inception embedding [113] of real and generated images. Different from FID, KID does not assume the distribution of activations to be parametric, and it also compares skewness as an addition to mean and variance.

**Learned Perceptual Image Path Similarity (LPIPS)** [116] measures the perceptual similarity between two images based on their deep embeddings computed by a pre-trained VGG model [117].

2) **Accuracy of Attribute Manipulation**: TARR and NAPR are the two most popular metrics for evaluating the overall accuracy of FAM methods.

**Target Attribute Recognition Rate** (TARR) is one of the most commonly used metrics to analyze whether the target attributes are presented in FAM results. It is usually measured by the proportion of images with the prediction of target attributes as the desired ones [33], [34], [40], [60], [65], [66], [69], [124], [137], [139], [143].

**Non-Target Attribute Preservation Rate** (NAPR) is the counterpart of TARR, which evaluates whether non-target attributes are preserved in FAM results [120]. It is usually measured by the ratio of non-target attributes being preserved [26], [37], [153], [154], or the degree of change in non-target attributes measured by certain estimators [155].

3) **Consistency of Image Semantics**: Consistency of image semantics can be measured by assessing the deviation caused by FAM methods across various aspects, such as pixel values, facial landmark coordinates, and head pose angles, and the commonly used metrics include the L1 norm, mean squared error (MSE), and root-mean-square error (RMSE).
TABLE I
CHARACTERISTICS OF FAM METHODS BASED ON IMAGE DOMAIN TRANSLATION

| Method Name | Publication | Model | Stability | Complexity | Latency | Quantities Metric |
|-------------|-------------|-------|-----------|------------|---------|------------------|
| CybeGAN [10] | ICCV 2017   | Cpn.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| DeepGAN [106] | ICCV 2017   | Cpn.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| GANsGAN [118] | ICCV 2017   | Cpn.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| UNIT [120]  | NeurIPS 2017 | Cpn.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| MUNIT [121]  | ICCV 2017   | Cpn.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| 3/settings [122] | CVPR 2017   | Cpn.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| AmortizedGAN [123] | TNNLS 2021 | Cpn.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| PGGAN [125]  | CVPR 2017   | GAN.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| Huang [126]  | arXiv 2019  | GAN.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| MCR [130]    | IJCAI 2019  | GAN.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| GANsGAN [132] | CVPR 2019   | GAN.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| ContextGAN [137] | CVPR 2020  | GAN.  | Two-domain | U/M         | 20s 1° 104 | LEI, USD |
| Context Neck [138] | ICLR 2020  | GAN.  | Multi-domain | U/M 20s 120 | LEI, USD |
| 3G-LS [139]  | ACM MM 2019 | GAN.  | Multi-domain | U/M 20s 120 | LEI, USD |
| MaskGAN [140] | CVPR 2019   | GAN.  | Multi-domain | U/M 20s 120 | LEI, USD |
| SegNet [94]  | arXiv 2018  | GAN.  | Absolute Label | U/M 20s 120 | LEI, USD |
| Tader Network [131] | NeurIPS 2017 | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| DPM [132]    | NeurIPS 2017 | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| PFCAN [133]  | ICCV 2018   | Cpn.  | Absolute Label | U/M 20s 120 | LEI, USD |
| TDCG [141]   | CVPR 2018   | Cpn.  | Absolute Label | U/M 20s 120 | LEI, USD |
| MemGAN [156] | CVPR 2019   | GAN.  | Absolute Label | U/M 20s 120 | LEI, USD |
| AxCG [33]    | TP 26 [19]  | GAN.  | Absolute Label | U/M 20s 120 | LEI, USD |
| UDPM [158]   | CVPR 2019   | GAN.  | Absolute Label | U/M 20s 120 | LEI, USD |
| SC [170]     | ACM MM 2019 | GAN.  | Absolute Label | U/M 20s 120 | LEI, USD |
| SMI [118]    | IC汇 2019   | GAN.  | Absolute Label | U/M 20s 120 | LEI, USD |
| HOCSGAN [116] | TP 26 [20]  | GAN.  | Absolute Label | U/M 20s 120 | LEI, USD |
| INT [140]    | ICCV 2020   | GAN.  | Absolute Label | U/M 20s 120 | LEI, USD |
| StarGANv2 [370] | CVPR 2021 | Cpn.  | Multi-domain | Style Code | U/M 20s 120 | LEI, USD |
| HEE [70]     | CVPR 2021   | GAN.  | Multi-domain | Style Code | U/M 20s 120 | LEI, USD |
| RankGAN [81] | ICCV 2019   | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| PPfGAN [84]  | ICCV 2019   | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| 3render [143] | ICCV 2020   | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| CoGAN [110]  | ICCV 2020   | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| GCGRN [148]  | ICCV 2020   | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| GANf [149]   | CVPR 2021   | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| LfAM [150]   | ICCV 2020   | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| HAO [151]    | IJCAI 2019  | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| PUNIT [180]  | ICCV 2019   | GAN.  | Multi-domain | Absolute Label | U/M 20s 120 | LEI, USD |
| HfGAN [152]  | Reference-based | Reference Image | U/M 20s 120 | LEI, USD |
| NeuralShop [154] | TP 26 [21] | User-template-based | U/M 122 122 | LEI, USD |
| t-NeuralShop [156] | TP 26 [22] | User-template-based | U/M 122 122 | LEI, USD |
| 3G-LS (312)  | ICCV 2019   | GAN.  | Multi-domain | U/M 20s 120 | LEI, USD |
| MemGAN [152] | CVPR 2019   | GAN.  | Multi-domain | U/M 20s 120 | LEI, USD |
| MemGAN [152] | CVPR 2019   | GAN.  | Multi-domain | U/M 20s 120 | LEI, USD |

Note: Cpn.: Cyclic; U/M: UMAP; M: Multi-setting; USD: User Study; LEI: Latent Editing Consistency; USD: User Study; LEI: Latent Editing Consistency.

Additionally, other metrics such as identity preservation, PSNR, and SSIM are frequently employed to evaluate the similarity between input and output images. Identity Preservation (IDPre) evaluates whether the identity of input faces is preserved in FAM results. It is usually measured by the face verification score [124], face recognition accuracy [40], distance/cosine similarity between identity features extracted by various networks [164, 165, 166].

**Peak Signal-to-Noise Ratio (PSNR)** [167] is commonly used to assess image degradation in the reconstruction process when the target attributes are set to be the same as the original ones (e.g., identity mapping).

**Structural Similarity Index Measure (SSIM)** [167] measures the structural similarity between two images in terms of illuminance, contrast, and structure. In addition, the Multi-scale SSIM (MS-SSIM) [168] is a variant of SSIM where the structural similarity is computed over multiple scales to incorporate image details at different resolutions.

4) **Properties of Latent Spaces**: The performance of FAM methods based on latent space navigation depends significantly on the characteristics of pre-trained generators measured by PPL, DSI, and LEC.

**Perceptual Path Length (PPL)** [30] measures the relationship between the scale of image changes and the step length of linear interpolation in latent spaces. **DCI [169]** quantitatively evaluates the disentanglement, completeness, and informativeness of latent representations when the ground-truth structure is available.

**Latent Editing Consistency (LEC)** [170] measures the extent to which the inaccuracy of GAN inversion translates to the error in subsequent editing results.

5) **Qualitative Measures**: In addition to quantitative measurements, a **User Study (US)** is also an important approach to evaluate the performance of FAM methods from the perspective of human perception [34, 66, 67, 72, 123, 144, 171, 172]. Specifically, participants are asked to choose one from multiple FAM results obtained by different candidate methods which they consider suits a certain criterion the best. The criteria can be flexibly designed to compare benchmark FAM methods in various dimensions.

IV. IMAGE TRANSLATION

Numerous methods consider FAM as an image-to-image translation (I2I) problem, where face images are grouped into visually distinct domains by their attributes. According to the number of image domains involved, this class of methods can be generally divided into two-domain and multi-domains methods.

A summary of image domain translation-based FAM methods is provided in Table I.

A. Two-Domain Methods

This class of methods aims at manipulating one facial attribute between two image domains X and Y. As one of the earliest studies, Pix2pix [62] uses an ordinary GAN-based model with the encoder-decoder architecture for image translation, and subsequent studies improve the resolution [173] and diversity [174].
of results. Although realistic translation results can be obtained, paired images \((x, y)\) are required to compute the pixel-wise loss for supervising the training process, which are usually very expensive to collect for many facial attributes (e.g., Age and Gender).

Therefore, FAM can also be considered as an unsupervised image translation problem where the main objective is to regularize the learning process based on unaligned face images with no ground truth. The cycle-consistency loss proposed in CycleGAN [63] (also in DiscoGAN [119], DualGAN [118], and UNIT [120]) is one of the most widely used constraints in unsupervised I2I approaches, where \(\mathcal{G} : X \rightarrow Y\) is coupled with an inverse mapping \(\mathcal{F} : Y \rightarrow X\) to compute the cyclic reconstruction image \(x' = \mathcal{F}(\mathcal{G}(x))\).

Numerous subsequent studies have been conducted to improve CycleGAN and the cycle-consistency loss for I2I tasks. AugCycleGAN [121] proposes to incorporate auxiliary random variables for modeling data variation and generating diverse results. To improve the spatial consistency, both ResGAN [64] and AttentionGAN [123] model FAM results as the combination of input images \(x\) and translation results \(\mathcal{G}(x)\). However, the pixel-level cycle-consistency constraint may not effectively deal with large geometric deformations in FAM (e.g., changing the hairstyle or removing eyeglasses). To solve this problem, ACGAN [122] introduces an additional discriminator to distinguish \(x\) between \(x^*\), which is adversarially trained with the translation network to enforce cycle-consistency from the data distribution perspective. In addition, CUT [126] proposes to incorporate contrastive constraints to preserve the source image’s structure when manipulating appearance.

Other approaches aim to solve the unsupervised FAM problem with one-sided (i.e., non-cyclic) models, which reduce the cost of training a reverse mapping function. Additional loss functions are usually incorporated to regulate the under-constrained mapping \(\mathcal{G}\). DIAT [124] enforces the high-level consistency between \(x\) and \(\mathcal{G}(x)\) with the perceptual loss and identity loss. MPCR [125] introduces an intermediate image domain \(X'\), and enforces the consistency between the translation results through two paths, i.e., \(X \rightarrow Y\) and \(X \rightarrow X' \rightarrow Y\). In addition, CouncilGAN [127] models \(\mathcal{G}\) with multiple translation networks and discriminators, which enables the generation of diverse FAM results.

\[ l \in \{1, 2, \ldots, N\} \]

### B. Multi-Domain Methods

To achieve multi-domain FAM with a single framework, one straightforward solution is to aggregate multiple two-domain networks into a single model [128], [129], [130]. However, the computational load would heavily increase as more facial attributes are considered. Thus, numerous approaches perform multi-domain FAM with a single pair of encoder and generator, where target attributes are indicated using label vectors. Qualitative and quantitative results of representative methods are shown in Fig. 5 and Table II.

The label vector, denoted as \(l = [l_1, l_2, \ldots, l_N] \in \mathbb{R}^N\), where \(l_i\) (\(i = 1, 2, \ldots, N\)) specifies the state of the \(i\)-th attribute, is one of the most widely used representations of facial attributes in the multi-domain FAM methods due to its flexibility and interpretability. According to the definition of \(l_i\), label vectors can be further divided into two types, i.e., absolute label vectors and relative label vectors.

#### 1) Absolute Label Vectors

Absolute label vectors describe the actual state of attributes for face images. As one of the earliest attempts, ICgan [60] uses label vectors with binary values as the conditional information, i.e., \(l = \{0, 1\}^N\), where \(l_i = 1\) \((l_i = 0)\) indicates the presence (absence) of the \(i\)-th facial attribute (see Fig. 6(a)). Specifically, one encoder network \(E\) embeds the input image \(x\) from any domain into the latent space, and one decoding network \(G\) generates the FAM result according to the target attribute vector \(l_y\), i.e., \(y = G(E(x), l_y)\).

To improve the consistency between input images and FAM results, AttGAN [33] introduces an identity mapping loss \(L_{idt} = \|G(E(x), I_x) - x\|\) to better preserve the image content in \(x\).
irrelevant to the target attributes, where \( G(E(x), l_y) \) refers to the identity mapping result of \( x \) conditioned on the label vector of original attributes \( l_y \) (denoted as \( x^r \) in Fig. 6(a)). Moreover, making use of a classification network \( C \), a cross-entropy-based attribute classification loss \( L_{cls} \) is also adopted to ensure the target attribute is modified properly. Using AttGAN as the backbone, HQGAN [139] achieves FAM on high-resolution face images (512 × 512) with the aid of a wavelet-based perceptual loss for improving the high-frequency textural details in translation results.

Similar to CycleGAN, StarGAN [34] exploits a cycle-consistency loss \( L_{cyc} = \|G(E(y), l_y) - x^r\|_1 \) for supervision, where \( G(E(y), l_y) \) is the cycle-reconstruction of \( x \). An attribute classification loss \( L_{cls} \) is used in StarGAN, which is implemented by multi-tasking the discriminator \( D \) instead of using an additional network as in AttGAN. Moreover, StarGAN augments the label vector with masks, which facilitates training and inference processes with multiple datasets containing different attribute labels. FPGAN [133] improves StarGAN by introducing the identity mapping loss as in AttGAN to the translation in both directions, which helps reduce unnecessary image modifications. In addition, SaGAN [134] introduces the spatial attention mechanism to localize image changes as in [123], where the attention map is dependent on the target attribute label \( l_y \). On the other hand, AttCycleGAN [135] adapts StarGAN to a face image super-resolution task where label vectors are used to control the attribute of generated HR images. For high-resolution face images, LATS [103] and HRFAE [147] propose to incorporate pre-trained StyleGAN [30] generators for synthesis.

The disentanglement between the latent embedding of \( z \) (i.e., \( z = E(x) \)) and the semantic information contained in \( l_y \) has also been studied. Both Fader Network [131] and UFDN [132] use an auxiliary latent discriminator \( D_z \) to identify the true label of \( z \) given \( z \) while the encoder \( E \) is trained to confuse \( D_z \). This adversarial learning process ensures that \( z \) is attribute-invariant, and thus the semantic of FAM results (i.e., \( y = G(z, l_y) \)) is guaranteed to be determined only by the information encoded in \( l_y \) and the influence of attributes in \( z \) is minimized.

To improve the accuracy of facial attribute translations with large geometric deformation, ADSPM [136] proposes a spontaneous motion estimation module to model the motion field for rendering attribute-driven shape transformation, e.g., change of expressions or shape of facial components, via an explicit warping operation. To deal with large changes in facial geometry, INIT [140] proposes to improve the visual quality of FAM results with large domain variations. Specifically, an adversarial importance weighting technique is designed to select more informative samples for training, and a multi-hop sampling strategy is used to stabilize the training process.

For diverse translation results, SDIT [137] integrates a random latent code \( z \) sampled from a Gaussian prior into the generator, such that re-sampling \( z \) at test time would produce multi-modal FAM results (as shown in Fig. 7(b)). Similarly, SMIT [138] proposes a domain embedding module to compute the guiding signal for FAM based on the target label vector \( l_y \) and a random style code \( z \sim \mathcal{N}(0, 1) \). To further improve the controllability of multi-modal FAM results, an encoder network \( E_s \) is incorporated in StarGANv2 [72] for computing the attribute-specific style representation of exemplar images, which is used to guide the generation of FAM results. Although diverse FAM results with controlled semantics can be obtained, irrelevant facial attributes (e.g., age and hair color) have also been modified, as shown in Fig. 7(b).

Thus, to improve the disentanglement between different facial attributes, HiSD [73] improves StarGANv2 by organizing facial attributes into a hierarchical structure with disentangled style allocation, and each attribute is associated with a dedicated translator network which produces diverse and localized modifications. FUNIT [148] achieves reference-based image translation in a few-shot setting, where exemplar images in a novel domain are utilized to control the semantics of synthesized results.

2) Relative Label Vectors: Several FAM methods use label vectors to indicate the difference in facial attributes between input and desired output images, i.e., relative label vectors. RelGAN [65] shows that the main problem of absolute label vector is that it requires specifying the entire set of attributes, which is inefficient since in practice most of them remain unchanged during FAM. To address this problem, RelGAN exploits the relative label vector, i.e., \( l_{rel} = l_y - l_x \), to inform the model what attributes to modify, and \( -l_{rel} \) naturally describes the reverse mapping process (see Fig. 6(b)). To ensure that the attribute difference between \( x \) and the FAM result \( y \) matches the relative label \( l_{rel} \), instead of using the classification loss \( L_{cls} \) as in AttGAN or StarGAN, an auxiliary discriminator is adversarially trained to distinguish between real triplets \((x, l_{rel}, x')\) (\( x' \) is a real image with the target attribute) and fake ones \((x, l_{rel}, y)\).

The relative label vector is also used in other FAM studies as conditional information. To help preserve the irrelevant content in input images without weakening the ability of attribute translation, SSCGAN [141] uses skip connections to transfer style features and spatial information in input images. In addition, skip connections are also incorporated in STGAN [66] and HifaGAN [145] to facilitate the forwarding of textural information in input images to FAM results, which are implemented with selective transfer units and wavelet-transformation blocks. In [142], CooGAN proposes a dual-path framework to perform FAM on HR images, where one branch focuses on generating fine-grained facial patches and the other monitors the overall facial structure. On the other hand, GCN-reprs [143] uses Graph Convolutional Networks to extract more structured information...
TABLE III
CHARACTERISTICS OF FACIAL SEMANTIC DECOMPOSITION-BASED FAM METHODS

| Method Name | Embedding | Description Level | Disentanglement | Max. Res. | Quantitative Metric |
|-------------|-----------|------------------|----------------|----------|--------------------|
| MUNIT [50] | Cyclic | Binary-dominant | Cyclic Consistency Loss | 215 × 215 | Conditions I–IV, LPSF, US |
| dCAN [157] | Cyclic | Binary-dominant | Cyclic Consistency Loss | 215 × 215 | TARR, US, Female Skin |
| cGAN [190] | Cyclic | Binary-dominant | Cyclic Consistency Loss | 215 × 215 | TARR, US, Male Skin |
| LGAN [179] | Cyclic | Binary-dominant | Cyclic Consistency Loss | 215 × 215 | TARR, US, Neutral Skin |
| LSGAN [208] | Cyclic | Binary-dominant | Adversarial Training | 215 × 215 | LPSF, TARR, Facial Expression |
| GFAN [207] | Multi-domain | Binary-dominant | Adversarial Training | 215 × 215 | LPSF, TARR, Facial Expression |
| CIGAN [307] | Pre-trained | Binary-dominant | Adversarial Training | 215 × 215 | LPSF, TARR, Facial Expression |
| LGAN [179] | Cyclic | Binary-dominant | Luminance Optimization | 215 × 215 | TARR, US, Neutral Skin |

Fig. 8. Illustration of (a) domain-level latent decomposition (X: smiling female subjects with blonde hair, Y: male subjects with black hair and neutral expression), and (b) instance-level latent decomposition with three facial semantics considered.

of high-level semantics and the relative dependencies of facial attributes, which is used to guide both the generator and discriminator for enhanced FAM results. In [144], WarpGAN further extends ADSPM [136] to manipulate HR face images with relative attributes, which is shown to be beneficial for generating localized image edits.

V. SEMANTIC DECOMPOSITION

Semantic decomposition approaches encode the input image \( x \) into multiple separate latent spaces, where the embedded representations are used to control different facial features. Thus, FAM can be performed by editing the latent code associated with target attributes. Based on the granularity of problem formulation, facial semantic decomposition can be further divided into two categories, i.e., domain-level decomposition and instance-level decomposition:

- **Domain-level decomposition** assumes that each image, regarded as a domain member, can be encoded into two separate latent spaces, one for modeling the domain-invariant information (i.e., Content Space \( C \)), and the other for capturing the domain-specific variations (i.e., Attribute Space \( A \)), as shown in Fig. 8(a).

- **Instance-level decomposition** focuses on factorizing facial semantics of individual input images into separate latent components (e.g., \( X_1, X_2, \ldots \) in Fig. 8(b)). This enables more flexible and disentangled control of facial attributes in finer granularity compared to learning domain-level translation patterns.

A summary of facial semantic decomposition-based FAM methods is provided in Table III. Fig. 9 shows qualitative results and comparisons of representative approaches in this category, and FID scores obtained from the respective original paper are reported in Table IV.

A. Domain-Level Decomposition

Based on the content space \( C \) and attribute space \( A \), the latent embeddings of an image \( x \in X \), denoted as \( c_x \) and \( a_x \), represent the semantic component that should be preserved and to be
manipulated. Given an exemplar image $y$ in the target domain $Y$, the translation of $x$ (denoted as $x'$) can be obtained based on its content embedding $c_x$ and the attribute code of $y$ (i.e., $a_y$), and $y$ can be similarly translated into $y' \in X$.

To learn a content-attribute decomposition between $C$ and $A_X/A_Y$, MUNIT [68] uses an image reconstruction loss to ensure that all information in $x$ is contained in $c_x$ and $a_x$. Moreover, a latent regression loss is incorporated to guarantee that encoders and generators are inverses of each other, which also implicitly facilitates the disentanglement between $c_x$ and $a_x$. On the other hand, DRIT [67] adopts an auxiliary content discriminator $D_C$ to distinguish the domain membership of $c_x$ and $c_y$, and the encoder networks are trained to confuse $D_C$ by making $c_x$ and $c_y$ only contain the domain-variant information. Leveraging such disentanglement, a cross-cycle consistency loss is proposed to regulate the learned mappings. These constraints are also proposed in cd-GAN [175] from the perspective of dual learning. In [69], a model similar to DRIT is developed where a group-wise whitening-and-coloring transformation is exploited to combine the embedding of content and attributes, which improves the memory and time efficiency as well as the quality of generation results.

For the multi-domain setting, numerous approaches model images from different domains with a common auto-encoder network, which is similar to the methods discussed in Section IV-B. DRIT++ [71] extends DRIT [67] to receive one-hot domain code at the input of attribute encoders and generators for specifying the source and target domain, and the disentanglement is achieved by adversarially training the encoders against an identity classifier (i.e., domain discriminator as in [67], [71]). Without using adversarial learning, LORD [178] learns a disentangled representation via latent optimization, and OverLORD [179] further analyzes the correlation between labeled and unlabeled facial semantics. In addition, LSM [180]...
solves a similar problem with the aid of the highly disentangled latent space of a pre-trained StyleGAN generator (see sample results in Fig. 9(c)). In contrast, SwapAE [70] preserves the overall structure of a face image, i.e., the shape and layout of facial components, and manipulates texture information according to exemplar images. Barbershop [181] also performs structure-texture decomposition by leveraging the latent space of a pre-trained StyleGAN2 [31] generator and is specifically designed to manipulate the color and style of hair.

Instead of low-dimensional latent codes, SensoriIn [186] operates on the geometry of a face image, which is described by various interpretable 2D representations inherently disentangled with facial textures (e.g., the feature map of facial landmarks), for synthesis. SoFGAN [187] proposes a semantic occupancy field to render parsed maps with arbitrary viewpoints, which are responsible for controlling the geometry of synthesized images.

Another line of research work disentangles content and attribute by introducing inductive bias into the structure of style-based generators. SNI [188] propose to condition input tensor c on another latent code \( z_c \) spatial content manipulation, which is disentangled against the style information controlled by the original latent code \( z \). In [189], DAT further makes \( z_c \) have a symmetric structure similar to \( z \) and controls the feature map at each scale. Recently, TransEditor [190] incorporates transformer blocks to establish the interaction between two latent spaces, which improves the controllability and flexibility of FAM.

2) Multiple Semantic Decomposition: Numerous methods are developed to encode face images into multiple latent components responsible for controlling different facial attributes based on contrastive training batch, 3D graphics model, face parsing map, and unsupervised factorizations.

Contrastive Training Batch: Numerous methods aim to explicitly associate latent components with different facial attributes, where the disentanglement is achieved by contrasting image samples in a designed training batch. Given a pair of face images \( x \) and \( y \) with the opposite label for the \( i \)-th attribute, DNA-GAN [74] divides their latent embeddings into multiple segments, and proposes to associate the \( i \)-th segments with the \( i \)-th attribute via contrastive learning techniques. ELEGANT [61] further improves DNA-GAN by making the translation networks only estimate the residual images, so that non-target attributes could be better preserved. In addition, GAN-Control [75] also constructs training batches to contain pairs of latent vectors. Unlike DNA-GAN and ELEGANT, paired latent codes in GAN-Control share only the \( i \)-th component and differ in all the rest parts. Thus, only the \( i \)-th attribute of corresponding images should be the same, and all other semantics should be different, which is also enforced by a contrastive loss.

3D Graphics Models: Apart from carefully designed training batches with paired images, numerous approaches leverage the inherently independent rigging parameters in 3D graphics models for learning disentangled semantic control. DiscoFaceGAN [36] divides the latent space of a pre-trained StyleGAN generator \( G^* \) into five subspaces, where the latent embeddings (denoted as \( \{ z_i \}_{i=1}^{5} \)) are responsible for controlling the shape, expression, pose, illumination, and all other textual details of the generated face \( G^*(z) (z = [z_1, z_2, z_3, z_4, z_5]) \), respectively (see Fig. 9(d)). The disentanglement among \( \{ z_i \}_{i=1}^{5} \) is achieved by adopting a contrastive loss computed based proxy faces rendered by 3DMM, which limits the change of image content caused by modifying \( z_i \) to the \( i \)-th attribute. Similarly, GIF [191] uses dense feature maps obtained by 3DMM for controlling the facial geometry, while disentanglement is achieved by ensuring the consistency of facial texture across different poses and expressions for the same identity. In addition, VariTex [192] further models the appearance with a variational latent code, which can be divided into two halves, one for modeling the identity information and the other for the hair and mouth interior.

Instead of 3DMM, ConfigNet [37] creates a set of face images with known semantic parameters and finer textural details following [203], and trains a generator \( G \) on SynthFace to model the mapping between semantic parameters and face images. In addition to 3D parametric models, physical face rendering models can also be used for facial semantic factorization. NFENet [38] uses the Lambertian rendering model [204] to decompose face images into three physically-based disentangled components, i.e., shape (face normals), albedo, and lighting, which can serve as tuning parameters for facial semantics. Recently, MOST-GAN [193] adopts a similar framework as [38] and proposes an iterative algorithm for fine-grained manipulation of hair regions.

Face Parsing Maps: In addition to parametric models which intrinsically factorize face images based on high-level semantics, face parsing maps are also widely used to encode spatially disentangled image content with separate latent spaces. Moreover, since parsing maps contain pixel-wise annotations of facial components, they can be regarded as representations of image layouts, which naturally allow the flexible manipulation of local regions.

With the aid of parsing map \( p \), MaskGuidedGAN [39] decomposes an input image \( x \) into five local facial components (i.e., left eye, right eye, mouth, skin & nose, and hair), and models each one with a separate auto-encoder network. In the testing phase, the latent representation of these objects can be combined with an arbitrary user-edited parsing map \( p' \) to render the corresponding face image, where the geometry and style are determined by \( p' \) and \( x \), respectively. MaskGAN [40] simplifies MaskGuidedGAN by computing a joint representation for the entire face instead of separately for each facial component, and an Editing Behavior Simulated Training strategy is proposed to simulate the user editing behavior on parsing maps. In [194], Semantic-StyleGAN uses a set of local generators to achieve spatially disentangled and semantically compositional face generation and manipulation, where parsing maps are only used for pixel-wise constraints.

Aside from the manipulation of each facial component, parsing maps can also be used with exemplar images to achieve fine-grained manipulation of low-level textural details. Unlike MaskGuidedGAN, SEAN [41] uses a single encoder network to compute the style matrix (ST) for input image \( x \) via a self-reconstruction process, where each element in ST encodes the appearance information of an image region indicated by \( p \). In the inference stage, elements in ST can be replaced by the
TABLE V

FID Scores of Selected Latent Navigation-Based FAM Methods

| Method Name | Gabor-T2 | FRID | | Method Name | Gabor-T2 | FRID |
|-------------|----------|------|------|-------------|----------|------|
| StyleGAN2Distillation [205] | - | 0.03 | - | StyleGAN2Distillation [205] | - | 0.03 |
| ResStyleGAN [33] | 21.4 | 19.1 | - | ResStyleGAN [33] | 21.4 | 19.1 |
| StyleGANv2 [34] | 14.7 | 12.6 | - | StyleGANv2 [34] | 14.7 | 12.6 |
| StyleGAN2 [35] | 13.2 | 11.5 | - | StyleGAN2 [35] | 13.2 | 11.5 |
| StyleGANv1 [36] | 13.4 | 11.3 | - | StyleGANv1 [36] | 13.4 | 11.3 |

FRID refers to the dataset of symbols images using generation. Presented FID scores obtained from the original paper are reported in Table V.

Fig. 10. Representative latent navigation-based FAM methods.

A. Linear Interpolation

Linear interpolation-based methods leverage the disentanglement of $G^*$’s latent spaces and model the translation process via simple linear interpolation, i.e., $z' = z_0 + \alpha n_Z$, where $n_Z$ is the unit vector of the direction in $Z$ associated with the target facial attribute and $\alpha$ is the step length. The focus of these methods is finding the semantically meaningful direction $n_Z$ corresponding to target attributes, with or without supervision.

1) Supervised Approaches: Supervised FAM approaches typically assume that the traversal direction $n_Z$ is uniquely determined by the target attributes and property of $G^*$’s latent space but independent of the input latent code $z_0$. Thus, $n_Z$ can be solved solely based on a large set of synthetic data, which is easy to collect due to the ability of $G^*$ in generating realistic images, and then applied to $z_0$ to obtain FAM results.

StyleGAN2Distillation [205] generates a synthetic face dataset by randomly sampling latent codes in $Z$, and considers $n_{W^i,j}$, the difference between the class center of the $i$-th and $j$-th attribute, as the corresponding traversal direction in $W^i$. The translation vector in StyleSpaceAnalysis [87] is computed in a similar way, but in the $S$ space for better disentanglement. ACU [206] extends StyleSpaceAnalysis by also manipulating feature maps in the generator, which can achieve more realistic FAM results without damaging the spatial disentanglement of image changes. On the other hand, InterFaceGAN [35] models $n_Z$ as the normal vector of hyper-planes separating latent codes corresponding to different attribute labels, and improves the disentanglement between two traversal directions via orthogonalization. Instead of annotating a synthetic dataset, GAN steerability [211] proposes to solve for latent traversing paths (linear or non-linear) in a self-supervised manner. StyleMapGAN [216] modifies the latent code of the StyleGAN generator to have spatial dimensions, which allows reference-guided facial component editing by blending such style maps. TargetCLIP [161] aims to solve the high-level semantic transfer problem (named essence transfer) with the guidance provided by CLIP [227].

Apart from attribute labels obtained by pre-trained classifiers, various annotation schemes are proposed to supervise the effect of latent navigation. To manipulate novel facial attributes (e.g.,

1For StyleGAN models [30, 31, 32], we use the $Z$ space for demonstrating operations in the latent space, and similar operations apply for the $W$, $W^+$, and $S$ spaces. The notation of equivalent concepts in different spaces share the same formulation and are distinguished by the specific symbol used, e.g., $z_0/w_0/w_0^+/s_0$ refers to the GAN inversion result in the $Z/W/W^+/S$ space, respectively. For other GAN models, e.g., SNGAN [226] (Spectral Norm GAN), BigGAN [80] and PGGAN [29], only the input latent space (denoted as $Z$) is analyzed.
TABLE VI
CHARACTERISTICS OF LATENT SPACE NAVIGATION-BASED FAM METHODS

| Method Name          | Publication     | Navigation | Latent Space | Max. Res. | Quantitative Metric                          |
|----------------------|-----------------|------------|--------------|-----------|---------------------------------------------|
| GAN user-orientation | ICLR 2020       | L.F./N.F.  | $Z_{adv}, W$ | 1024 x 1024 | FID, LIPS, Fréchet’s correlation coefficient |
| StyleGAN2Distillation | ECCV 2020       | L.1        | $W, W^{n}$  | 1024 x 1024 | FID, US                                      |
| GenFacGAN [35]       | CVPR 2020       | L.1        | $Z_{adv}, W$| 1024 x 1024 | Correlation of Attribute Distributions      |
| ACU [65]             | ACM MM 2021     | L.1        | $S$          | 1024 x 1024 | FID, AD [67], Beta score of Local Editing, Region Purity |
| AdvStyle [74]        | CVPR 2020       | L.1        | $W$          | 1024 x 1024 | Correlation of Attribute Distributions      |
| EditingGAN [159]     | NeurIPS 2021    | L.1        | $W$          | 1024 x 1024 | FID, XID, TARR, LIPS                       |
| EditGAN [159]        | ICLR 2021       | L.1        | $W, W'$     | 1024 x 1024 | NAIR, IPDS, US                             |
| Label-Transformation [24] | ICCV 2021   | L.1        | $W, W'$     | 1024 x 1024 | FID, LIPS, SWD, AD [67], Cost Analysis    |
| Style-Transformer [107] | CVPR 2020 | L.1        | $W$          | 1024 x 1024 | TARR, LIPS                                  |
| UDEH [32]            | ICM 2020        | L.1        | $Z_{adv}, W$| 1024 x 1024 | TARR, US                                   |
| WarpEditingGAN [213] | ICCV 2021       | L.1        | $Z_{adv, W}$| 1024 x 1024 | TARR, L-normalized Correlation of Attribute Distributions |
| GANSpace [214]       | NeurIPS 2020    | L.1        | $Z_{adv}, W$| 1024 x 1024 | -                                           |
| SfA [79]             | CVPR 2020       | L.1        | $Z_{adv}, W$| 1024 x 1024 | FID, US, Attribute Re-scoreing Analysis    |
| LowRankGAN [106]     | NeurIPS 2021    | L.1        | $Z_{adv, W}$| 1024 x 1024 | FID, IPDS, Masked L2 Error of Final Value, US |
| LatentCLB [115]      | ICCV 2021       | L.1        | $Z_{adv}, W$| 1024 x 1024 | FID, US, Attribute Re-scoreing Analysis    |
| StyleMapGAN [216]    | CVPR 2021       | L.1        | $W$          | 1024 x 1024 | FID, LIPS, MS, Inference Speed, Average Precision |
| AugGAN [161]         | ECCV 2021       | L.1        | $W$          | 1024 x 1024 | FID, IPDS, Visual Semantic Alignment       |
| SAM [217]            | TOG 2021        | L.1        | $W$          | 1024 x 1024 | US, Differences of the predicted target age |
| Hair-CLIP [143]      | CVPR 2020       | L.1        | $W$          | 1024 x 1024 | IPDS, PINN, SSD, US, Average Differences of Hair Color |
| CLIP-StyleGAN [218]  | NeurIPS 2021    | L.1        | $W$          | 1024 x 1024 | US, Semantic Semantic Alignment             |
| Brn-supervised [219] | CVPR 2022       | L.1        | $W$          | 1024 x 1024 | US, Control/Disentanglement Curve          |
| NoneGan [220]        | ICCV 2021       | N.1        | $W$          | 256 x 256  | US, Manipulation Disentanglement Curve/Score |
| SGP [221]            | CVPR 2021       | N.1        | $Z_{adv}, W$| 1024 x 1024 | FID, Inference Disentanglement Curve/Score   |
| StyleGAN2 [133]      | CVPR 2021       | N.1        | $Z_{adv}, W$| 1024 x 1024 | FID, LIPS                                    |
| StyleGAN2 [77]       | CVPR 2020       | N.1        | $Z_{adv}, W$| 1024 x 1024 | FID, LIPS, SSD, L2 Error of Attribute Edit Consistency |
| PIGE [118]           | COG 2020        | N.1        | $W$          | 1024 x 1024 | FID, IPDS, L2 Error of Attribute Edit Consistency |
| StyleGAN [79]        | ICCV 2020       | N.1        | $Z_{adv}$   | 1024 x 1024 | FID, TARR, IPDS, Disentangled Edit Strength, Inference Time |
| DeGAN2 [134]         | ICCV 2021       | N.1        | $W$          | 1024 x 1024 | FID, IPDS, SSD, Inference Disentanglement Curve/Score |
| SAMIC [145]          | ICCV 2021       | N.1        | $W$          | 1024 x 1024 | FID, IPDS, SSD, Inference Disentanglement Curve/Score |
| DeGAN-V2 [119]       | CVPR 2020       | N.1        | $W$          | 1024 x 1024 | FID, TARR, IPDS, US                        |
| LowRankGAN [216]     | CVPR 2021       | N.1        | $W$          | 1024 x 1024 | FID, IPDS, SSD, Inference Disentanglement Curve/Score |
| exquisiteGAN [217]   | CVPR 2021       | N.1        | $W$          | 1024 x 1024 | FID, LIPS, IPDS, SSD, Inference Disentanglement Curve/Score |
| StyleGAN [222]       | CVPR 2021       | N.1        | $W$          | 1024 x 1024 | FID, LIPS, SSD, Inference Disentanglement Curve/Score |
| ClipGAN2 [223]       | ICCV 2021       | N.1        | $W$          | 1024 x 1024 | FID, TARR, IPDS, US                        |
| Talk-to-edit [156]    | CVPR 2021       | N.1        | $W, W$      | 1024 x 1024 | IPDS, NAIR, US                             |
| AnyGAN [224]         | CVPR 2021       | N.1        | $Z, W$      | 1024 x 1024 | FID, LIPS, Relative Face Recognition Rate   |
| Sound-guided [125]   | CVPR 2021       | N.1        | $Z, W$      | 1024 x 1024 | US                                         |

Latent space directions in which latent space navigation is performed: source $W$, $W'$, and $Z_{adv}$ refer to latent spaces in adversarial generation, and $Z_{adv}$, $Z_{adv}$, and $Z_{adv}$ denote the latest spaces of (PHAN, POSEGAN, and GAN) respectively. For “Navigation”, L.F. and N.F. refer to “Linear interpolation” and “Non-linear manner”, respectively. For “Max. Res.”, $Z_{adv}$ and $Z_{adv}$ are the maximum resolution of $Z_{adv}$ and $Z_{adv}$, respectively. 

Fig. 11. Sample FAM results obtained by (a) AdvStyle [76] (input images are in the middle) and (b) EditGAN [159] (results obtained with latent editing directions learned on user-edited samples). Target facial attributes are labeled underneath images.

Supermodel and Chinese Celebrity, see Fig. 11(a)) beyond well-established ones with binary labels, AdvStyle [76] introduces a deep model to discriminate the target attribute from transformed images, which is simultaneously updated with $n_W$. To perform user-defined geometric changes of facial components (as shown in Fig. 11(b)), EditGAN [159] allows users to edit the parsed map of an input image and finds the latent displacement vector to reflect such modification via an optimization-based method.

However, some studies point out that the latent space of $G'$ is not guaranteed to be uniform, and thus $n_Z$ is not globally applicable and should be dependent on the source latent embedding, i.e., $n_Z = n_Z(z_0)$. EnjoyEditingGAN [155] uses a simple multilayer perceptron (MLP) to compute the translation vector based on $z_0$ for manipulating one facial attribute. As it learns multiple traversal directions for $W$ simultaneously, it is likely to disentangle and control image changes better. SAM [217] adopts a network module similar to the pSp encoder [228] for computing the displacement vector in $W$ space. Hair-CLIP [162] incorporates three sub-mapper networks to compute the displacement vector at different granularity levels (i.e., coarse, medium, and fine). Latent-Transformer [26] uses a transformer in $W$ to predict the traversing direction $n_{W'}$ based on the embedding of the input images $w_{0}^{*}$. In addition, transformer blocks are also used in Style-Transformer [207] to achieve flexible FAM based on both attribute labels and exemplar images.

2) Unsupervised Approaches: To reduce the annotation cost of training data, numerous unsupervised methods have been proposed to discover interpretable semantic directions $n_Z$ for FAM. The key is to describe the properties of the desired $n_Z$ with a concrete objective function, which can be solved via optimization.

Without any form of supervision, UDID [212] discovers a set of traversal directions where the induced image transformations could be distinguished from each other. Specifically, an auxiliary regression network $R$ is adopted to predict $n_Z$ and the $a$ from a pair of images $G(z)$ and $G(z + n_Z)$, which is optimized jointly with $n_Z$ in a self-supervised manner. WarpedGANSpace [213] adopts a framework similar to UDID with $n_Z$ estimated by a set of radial basis functions (RBFs). On the other hand, GANSpace [214] models interpretable traversal directions as the principal components of feature tensors in $W$, which capture the major semantic variations of training data (i.e.,
facial attributes) and can be solved by Principal Components Analysis. Recently, SeFa [79] considers the projection applied to the input latent code \( z \in \mathbb{Z} \) (denoted as \( G_1 \)) as an affine transformation, \( y = G_1(z) = Az + b \) (\( A \) and \( b \) are weight and bias parameters), and shows that important semantics changes should be associated with latent directions \( n_Z \) that maximize the variation of the output of \( G_1 \). In addition, LowRankGAN [208] models the relationship between \( z \) and \( G^*(z) \) using the Jacobian matrix, and solves \( n_Z \) by maximizing the change of image \( G(z) \). It also shows that interpolating along the determined direction \( n_Z \) can generate spatially disentangled FAM results. LatentCLR [215] assumes variation of visual appearance caused by modifying \( z \) could be identifiable by \( n_Z \). Thus, \( n_Z \) can be obtained by imposing contrastive constraints. In addition, CLIP2StyleGAN [218] integrates the pre-trained latent spaces of StyleGAN and CLIP [227], to extract semantically-labeled FAM directions.

**B. Non-Linear Traversal**

There are also many methods that assume the latent spaces of \( G^* \) are non-linear and propose to make \( z_0 \) travel along a complex non-linear trajectory \( T \). Thus, \( z' \) can be written as \( z' = T(z_0) \) where \( T \) is usually implemented by iterative optimization or deep neural networks. The trajectories of non-linear traversal can mainly be computed with iterative optimization or deep neural networks.

1) **Iterative Optimization:** Several methods operate on a non-linear latent space by dividing the traversing process into multiple steps, and interpolating within a neighborhood at each step. NeuralODE [220] parameterizes the traversal in latent space by Neural ODE [229], and thus linear interpolation could be considered as a special case where the first derivative is a constant, i.e., \( \dot{w} = n_W \) (the initial condition is set to \( w_0 \)). Following this formulation, non-linear latent traversal could be realized by replacing the right-hand side term with \( f(w) \), where \( f \) is usually implemented with deep neural networks. SGF [221] specifies \( f \) as the surrogate gradient field determined by the desired property of manipulation results (e.g., facial attributes or landmark layout). In [78], StyleFlow utilizes a bi-directional continuous normalizing flow (CNF) based network that is conditioned on attribute labels to model the mapping between latent codes in \( Z \) and \( W \) space. SSFlow [160] also uses a flow-based framework for achieving FAM, and makes several improvements to the network structure to enhance the identity-preserving ability.

To approximate non-linear latent trajectories, HijackGAN [153] trains a proxy model \( P \) to map \( z \) to \( M \circ G^*(z) \) (\( M \) is an attribute classifier), and thus the \( j \)-th row vector of \( P \)’s Jacobian matrix describes the traversal direction for manipulating the \( j \)-th attribute at each step. In the iterative update process proposed in IALS [172], the traversal direction at each step contains two parts: an instance-specific semantic direction \( d_l \) (i.e., local direction dependent on \( z \)) and an attribute-level semantic direction \( d_g \) (i.e., global direction independent of \( z \)). A theoretical understanding of the geometric structure of latent spaces in GANs can be found in [230], [231].

2) **Deep Neural Networks:** Apart from iterative interpolation-based methods, deep neural networks are also used to determine the target latent code \( z' \). Given the latent embeddings of source and exemplar face images (denoted as \( w_0 \) and \( w_e \), respectively), StyleRig [77] computes the rigging parameter \( \Delta w \) as \( \Delta w = D(E_0(w_0), E_e(w_e)) \), where \( E_0 \), \( E_e \), and \( D \) are implemented using MLP networks. PIE [158] improves StyleRig by jointly optimizing the image embedding and manipulation modules. LACE-ODE [209] represents the joint distribution of data and attributes with an energy-based model (EBM) in the latent space, and thus FAM could be performed by changing the combination of latent attribute codes. GuidedStyle [163] proposes to use attention-based residual MLP blocks to map latent code in \( W \) to \( W^+ \). DyStyle [210] designs an attribute-conditioned dynamic network for manipulating the latent code in \( W^+ \), and leverages contrastive learning to enhance the disentanglement of different attributes. GH-Feat [156] considers the pre-trained StyleGAN generator as a learned loss function, and trains an encoder network to compute hierarchical features in the \( S \) space.

**VII. PROPERTIES OF GAN-BASED FAM METHODS**

In this section, we summarize the important properties of GAN-based FAM methods in five aspects, including visual quality, desired target attribute, semantic consistency, diversity and controllability, and editing multiple features.

A. Visual Quality of FAM Results

FAM results are expected to be as photo-realistic as possible in terms of both the overall structure and fine-grained details. The improvement in visual quality and image resolution can largely be attributed to the rapid development of GAN-based image synthesis methods, including training objectives [27], [81], [82], [83] and network structures [28], [29], [30], [31], [32]. These advancements help stabilize the adversarial training process between the generator and discriminator, and eliminate ghost artifacts in synthesized images. The recent success of style-based GANs [30], [31], [32] in synthesizing HR images and learning disentangled semantic representations has enabled efficient FAM based on latent space navigation [35], [79], [214].

B. Desired Target Attributes

In FAM, the modified target attributes should be successfully recognized in synthesized images. This goal is usually achieved by explicitly imposing an attribute classification loss [33], [34] or implicitly associating different latent components with disentangled facial attributes [41], [75].

Attribute classification loss is usually used in FAM methods based on image domain translation and latent space navigation. FAM models learn to render facial attributes as desired in translation results by minimizing the attribute classification loss, which is commonly implemented by using an auxiliary recognition network [33], [76], [139] or multi-tasking the discriminator [34], [66], [137].

For semantic decomposition-based FAM methods, desired results are obtained by manipulating the latent components
corresponding to target attributes. Thus, the key is to closely link each decomposed facial semantic to a designated latent component, which is usually ensured by imposing contrastive losses [61], [74], [75] or involving parametric 3D graphics models [36], [37], [38], [191], [192], [193].

C. Preservation of Facial Semantics

In addition to correctly synthesizing images with desired attributes, other facial regions are also expected to be preserved after manipulation. This can be achieved by imposing direct constraints on image consistency or better disentanglement among semantic regions.

Pixel-level constraints are widely used to enforce modifications only on regions closely related to the target attributes, which are either learned during training [123], [134] (e.g., via the attention mechanism) or indicated by face parsing maps [232]. Some methods regulate the synthesis process by penalizing the difference of all pixels between input and reconstructed images, which can be obtained by cyclic translation [63], [118], [119] (i.e., cycle-consistency loss) or identity mapping [33], [133]. Moreover, feature-level constraints are also employed by FAM methods to ensure perceptual consistency between input and edited images (e.g., identity preservation [124]).

On the other hand, some methods propose to disentangle the factors for different facial attributes, and thus the manipulation of target attributes can be constrained to proper regions or semantics. Such disentanglement can be achieved by adversarially training with an auxiliary discriminator [67], [71], contrastive learning with specially designed training batches [61], [74], [75], leveraging the intrinsic disentanglement of 3D models [36], [37], [191], [192], or using face parsing maps for locating individual facial component [39], [40], [41], [194], [200], [201], [202].

D. Diversity and Controllability of Generated Images

Facial attributes typically describe perceptually salient properties of human faces, where no specific restrictions are imposed on the shapes or textures of facial components in manipulated images. Thus, FAM models should be able to generate diverse (also referred to as multi-modal) outputs. However, most image domain translation-based methods [33], [34], [63] are unimodal, and only inter-domain mapping functions are estimated. To address this issue, a few approaches [137], [138] incorporate random variables for modeling intra-domain variation of the target domain, and thus diverse FAM results can be obtained by re-sampling at test time.

To improve the controllability of multi-modal results, some FAM methods use style information [67], [71], [72], [73], which encodes the intra-domain variation specific to a certain subject (i.e., exemplar face), for guiding the texture and geometry of translation results. In particular, a few methods [67], [71] align the distribution of style code with random variables, which makes the model compatible with two types of data as guidance, i.e., noise vectors and exemplar images. Multi-modal FAM methods can also be obtained by simply style mixing [30], [31] with latent codes sampled from the prior distribution or extracted from exemplar images for better controllability.

E. Editing Multiple Attributes

For image domain translation-based FAM results, two-domain approaches [64], [123], [124] can only manipulate one facial attribute with a single model. The capacity of multi-domain approaches [128], [129], [130] is determined by the number of attribute labels available in training data, and the model has to be re-trained whenever new attributes are introduced for editing.

For semantic decomposition-based FAM methods, the number of features that can be modified is largely determined by the internal mechanism for dividing facial semantics. Specifically, FAM methods trained with contrastive data [61], [74], [75] can be extended to manipulate multiple features (as long as paired images with the target attribute can be collected), but may be less effective for disentanglement as the number of attributes increases. However, FAM methods based on parametric 3D models [36], [37], [191], [192] and face parsing maps [39], [40], [41] for explicit semantic decomposition cannot be easily extended to handle multiple features as specific prior knowledge need to be incorporated (e.g., the number of categories in parsing maps).

FAM methods based on latent space traversal can be easily extended to manipulate multiple features, as the generator pre-trained on unconditional image synthesis has learned to extract the representation of different semantics. Moreover, since the parameters of the generator are fixed, the cost of adapting this class of FAM methods to edit unseen attributes is lower than that of other approaches.

VIII. DISCUSSION

A. Challenges and Future Directions

Fine-grained Control of Individual Components: Most existing GAN-based FAM methods, conditioned either on label vectors or style codes, can hardly provide low-level controls on the shape and texture of manipulation results. For instance, fine-grained control of the exact shape and structure of hair, the angle of gaze direction, and head pose, cannot be accomplished when editing the associated attributes by existing approaches. To solve this problem, some studies incorporate pixel-wise conditional information obtained by user interaction [150], [151], [152], [159] as guidance.

Task-Specific Inductive Biases: Although general-purpose image translation methods (e.g., CycleGAN [63]) can be used for editing facial attributes, introducing task-specific inductive biases can help further improve the quality of FAM results. Most recently, Neural Radiance Fields (NeRFs) have been used as an implicit presentation to model fine-grained 3D structure of faces [183], [233], [234], [235], where the volumetric rendering mechanism is introduced for reconstruction. Thus, developing mechanisms to introduce more efficient inductive biases specific to the target facial attribute is a promising future research direction.

Biased Generative Priors: Although pre-trained generators are widely used in FAM methods for image synthesis, the distribution of output images is restricted and biased towards
the training dataset. For example, most face images synthesized by StyleGAN generators pre-trained on FFHQ have neutral or smiling expressions with the near-frontal view. Thus, it is difficult for these networks to model faces with large poses or arbitrary expressions [47]. It is of great interest to mitigate such biased priors by adapting the domain-agnostic semantics learned by pre-trained GAN models for tasks with different data distributions.

Guiding Information With Multimodality: In most existing GAN-based FAM methods, the target attributes are specified by labels [33], [34], [72] or exemplar images [41], [61], [73]. Recently, conditional information in other modalities, such text [154], [182], [223], [224], [236] and speech [237], [238], [239], has attracted increasing research attention due to the development of pre-trained large-scale frameworks (e.g., CLIP [227]) and availability of related datasets (CelebA-Dialog [240]). Moreover, novel modalities of supervision signal, such as biometrics (e.g., brain responses recorded via electroencephalography [219]) and sound [225], have also been utilized as guidance for semantic editing.

Video-Based FAM: The recent development in DeepFake [46], [241] and face re-enactment [242], [243] have demonstrated great societal impact and significant practical value of FAM on video data. Moreover, video manipulation can also serve as a data augmentation method for DeepFake detection methods, and thus plays an important role in information forensics. Most recently, a new video dataset with facial attribute annotation of emotion, action, and appearance is released (CelebV-HQ [240]), which shows growing interest in video-based FAM.

B. Other Issues Related to GAN-Based FAM

Alternative Taxonomy of FAM Methods: This paper groups FAM methods based on initial motivations rather than technical similarities or original purposes. While some methods such as MUNIT [68] and DRIT [67] were initially proposed for image translation, they are classified as semantic decomposition-based approaches as their fundamental motivation are similar to other methods in this category. There are alternatives to discuss these methods for specific research tasks or contexts.

Comparison Between Other Generative Models: Other generative models, including VAEs [244], flow-based models [245], and diffusion models [246], [247], can also be used for performing FAM in addition to GANs. VAE has a stable training process but tends to generate blurry and unrealistic images. Flow-based models can generate high-quality images but have higher computational complexity and less flexibility in model design than GAN models.

Diffusion models have recently demonstrated comparable or better performance in generating and manipulating diverse images, in contrast to GANs [248]. They can be applied to a wider range of images and achieve better trade-offs between input reconstruction and editability of inverted latent code, while enjoying a more stable training process and diverse generation results [249], [250], [251]. However, these methods have slower inference speeds than GAN-based models due to the multi-step de-noising process, and for text-guided FAM tasks, GAN-based models have been shown to achieve better language-vision alignment [251].

Influence of GAN Inversion Techniques: To use pre-trained unconditional generators on real images, it is necessary to first apply GAN inversion techniques [44] to convert real images into the latent space. One of the primary goals of GAN inversion techniques is to reconstruct input real images with perceptual fidelity accurately [85], [86], [228], [252]. However, other studies [170], [253] have also emphasized that the visual quality and semantic disentanglement of manipulation results in downstream tasks obtained by varying the latent code are equally important. This property, known as the editability of inverted latent codes, is often considered to involve a trade-off between the faithfulness of reconstruction [170]. In addition, some recent studies [254], [255] also use hypernetworks [256] to enhance the overall performance of GAN inversion techniques.

IX. CONCLUSION

Facial attributes intuitively describe the representative properties of human faces, and have received much attention in the field of vision and learning. As one of the most widely-used research topics, facial attribute manipulation has both important research and application values. This paper presents a comprehensive survey of existing GAN-based FAM studies. We analyze the similarities and differences between these methods in terms of motivation and technical details. Moreover, we summarize the important properties of FAM as well as how they are approached by existing methods. We conclude this survey with discussions of the challenges and future research directions of FAM.

REFERENCES

[1] G. Hu et al., “Attribute-enhanced face recognition with neural tensor fusion networks,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 3764–3773.
[2] C. Huang, Y. Li, C. C. Loy, and X. Tang, “Deep interleaved learning for face recognition and attribute prediction,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2781–2794, Nov. 2020.
[3] L. Fang, J. Zhang, and B. Deng, “Robust RGB-D face recognition using attribute-aware loss,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 10, pp. 2552–2566, Oct. 2020.
[4] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, “Heterogeneous face attribute estimation: A deep multi-task learning approach,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 11, 2597–2609, Nov. 2018.
[5] A. Zaeranzadeh et al., “Face image retrieval with attribute manipulation,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 12116–12125.
[6] Y. Wu et al., “Attribute prototype learning for interactive face retrieval,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 2593–2607, 2021.
[7] Y. Wu, N. M. Thalmann, and D. Thalmann, “A plastic-Visco-Elastic model for wrinkles in facial animation and skin aging,” in Proc. 2nd Pacific Conf. Fundam. Comput. Graph., 1994, pp. 201–213.
[8] D. A. Rowland and D. I. Perrett, “Manipulating facial appearance through shape and color,” IEEE Comput. Graph. Appl., vol. 15, no. 5, pp. 70–76, Sep. 1995.
[9] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3D faces,” in Proc. Ann. Conf. Comput. Graph. Interactive Techn., 1999, pp. 187–194.
[10] T. Hassner, “Viewing real-world faces in 3D,” in Proc. IEEE Int. Conf. Comput. Vis., 2013, pp. 3607–3614.
[11] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, “DeepFace: Closing the gap to human-level performance in face verification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 1701–1708.
[12] T. Hassner, S. Harel, E. Paz, and R. Enbar, “Effective face frontalization in unconstrained images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4295–4304.

[13] C. Cao, Q. Hou, and K. Zhou, “Displaced dynamic expression regression for real-time facial tracking and animation,” ACM Trans. Graph., vol. 33, no. 4, pp. 1–10, 2014.

[14] X. Zhu, Z. Lei, J. Yan, D. Yi, and S. Z. Li, “High-fidelity pose and expression normalization for face recognition in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 778–796.

[15] P. Ekman and W. V. Friesen, “Facial action coding system: A technique for the measurement of facial movement,” Palo Alto: Consulting Psychologists, 1978.

[16] B. Facel and J. Luettin, “Automatic facial expression analysis: A survey,” Pattern Recognit., vol. 36, no. 1, pp. 259–275, 2003.

[17] C. Shan, S. Gong, and P. W. McOwan, “Facial expression recognition based on local binary patterns: A comprehensive study,” Image Vis. Comput., vol. 27, no. 6, pp. 803–816, 2009.

[18] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, “Synthesizing realistic facial expressions from photographs,” in Proc. SIGGRAPH Courses, 2006, Art. no. 19.

[19] E. B. Roesch, L. Tamarit, L. Reveret, D. Grandjean, D. Sander, and K. R. Rother, “Synthesizing realistic facial expressions from photographs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 27, no. 11, pp. 3458–3465, 2020.

[20] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, “Face animation using a 3D facial model and motion capture data,” ACM Trans. Graph., vol. 25, no. 3, pp. 795–802, 2006.

[21] D. Vlasic, M. Brand, H. Pfister, and J. Popovic, “Face transfer with disentangled face editing in images and videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 787–796.

[22] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. B. Johnson, “DeepFakes and beyond: A survey of face manipulation and fake detection,” Inf. Fusion, vol. 64, pp. 131–148, 2020.

[23] T. Karras, S. Laine, and T. Aila, “A style-based generator architecture for generative adversarial networks,” in Proc. Neural Inf. Process. Syst. Conf., 2017, pp. 8789–8797.

[24] Z. Wan et al., “Old photo restoration via deep latent space translation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 299–311.

[25] M. Kowalski, S. J. Garbin, V. Estellers, T. Baltrusaitis, M. Johnson, and J. Shotton, “CONFIG: Controllable neural face image generation,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 299–311.

[26] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Comaniciu, “Neural face editing with intrinsically disentangled latent spaces,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 5541–5550.

[27] S. Gu, J. Bao, H. Yang, D. Chen, F. Wen, and L. Yuan, “Mask-guided portrait editing with conditional GANs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3456–3457.

[28] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, “MaskGAN: Towards diverse and interactive facial image manipulation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5549–5558.

[29] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, “SEAN: Image synthesis with semantic region-adaptive normalization,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5103–5112.

[30] X. Zheng, Y. Guo, H. Huang, Y. Li, and R. He, “A survey of deep facial attribute analysis,” Int. J. Comput. Vis., vol. 128, no. 8, pp. 2002–2034, 2020.

[31] Y. Pang, J. Lin, T. Qin, and Z. Chen, “Image-to-image translation: Methods and applications,” IEEE Trans. Multimedia, vol. 24, pp. 3859–3881, 2022.

[32] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, and M.-H. Yang, “GAN inversion: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 3, pp. 3121–3138, Mar. 2023.

[33] X. Wang, K. Wang, and S. Lian, “A survey on face data augmentation for the training of deep neural networks,” Neural Comput. Appl., vol. 32, no. 19, pp. 15503–15531, 2020.

[34] R. Tolosana, R. Vera-Rodríguez, J. Fierrez, A. Morales, and J. Ortega-García, “DeepFakes and beyond: A survey of face manipulation and fake detection,” Front. Inf. Fusion, vol. 64, pp. 131–148, 2020.

[35] Y. Zhu, Q. Li, J. Wang, C.-Z. Xu, and Z. Sun, “One shot face swapping on megapixels,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 4834–4844.

[36] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, “Information bottleneck disentanglement for identity swapping,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 3404–3413.

[37] Q. Gu, G. Wang, M. T. Chiu, Y.-W. Tai, and C.-K. Tang, “LADN: Local adversarial disentangling network for facial makeup and de-makeup,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 10481–10490.

[38] W. Jiang et al., “PSGAN: Pose and expression robust spatial-aware GAN for customizable makeup transfer,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5194–5202.

[39] M. Maximov, I. Elezi, and L. Leal-Taixé, “CIAGAN: Conditional identity anonymization generative adversarial networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5447–5456.

[40] B. Meden et al., “Privacy-enhancing face biometrics: A comprehensive survey,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 4147–4183, 2021.

[41] X. Yu, B. Fernando, R. Hartley, and F. Porikli, “Semantic face hallucination: Super-resolving very low-resolution face images with supplementary attributes,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 11, pp. 2926–2943, Nov. 2020.

[42] S. Menon, A. Damian, S. Hu, N. Ravi, and C. Rudin, “PULSE: Self-supervised photo upsampling via latent space exploration of generative models,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2437–2445.

[43] Z. Wan et al., “Old photo restoration via deep latent space translation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 2, pp. 2071–2087, Feb. 2023.

[44] J. Kim, M. Kim, H. Kang, and K. H. Lee, “U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation,” in Proc. Int. Conf. Learn. Representations, 2019.

[45] G. Song et al., “AgileGAN: Stylizing portraits by inversion-consistent transfer learning,” ACM Trans. Graph., vol. 40, no. 1, pp. 1–13, 2021.

[46] R. Tolosana, R. Vera-Rodríguez, J. Fierrez, A. Morales, and J. Ortega-García, “DeepFakes and beyond: A survey of face manipulation and fake detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 5, pp. 1025–1037, May 2020.

[47] J. Lee, E. Kim, Y. Lee, D. Kim, J. Chang, and J. Choo, “Reference-based sketch image colorization using augmented-self reference and dense semantic correspondence,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5801–5810.

[48] G. Perarnau, J. Van De Weijer, B. Raducanu, and J. M. Álvarez, “Invertible conditional GANs for image editing,” 2016, arXiv: 1611.06355.
L. Hui, X. Li, J. Chen, H. He, and J. Yang, “Unsupervised multi-domain image translation with domain-specific encoders/decoders,” in Proc. Int. Conf. Neural Inf. Process. Syst., vol. 2021, pp. 2595–2604.

M. M. R. Siddiquee et al., “Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 191–200.

G. Zhang, M. Kan, S. Shan, and X. Chen, “Generative adversarial network with spatial attention for face attribute editing,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 417–432.

Y. Lu, Y.-W. Tai, and C.-K. Tang, “Attribute-guided face generation using conditional CycleGAN,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 282–297.

R. Wu, X. Tao, X. Gu, X. Shen, and J. Jia, “Attribute-driven spontaneous motion in unpaired image translation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 5923–5932.

Y. Wang, A. Gonzalez-Garcia, J. van de Weijer, and L. Herranz, “SDIT: Scalable and diverse image translation,” in Proc. ACM Int. Conf. Multimedia, 2019, pp. 1267–1276.

A. Romero, P. Arbeláez, L. Van Gool, and R. Timofte, “SMIT: Stochastic multi-label image-to-image translation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops, 2019, pp. 3285–3294.

Q. Deng, Q. Li, J. Cao, Y. Liu, and Z. Sun, “Controllable multi-attribute editing of high-resolution face images,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 1410–1423, 2021.

J. Cao, H. Huang, Y. Li, R. He, and Z. Sun, “Informative sample mining network for multi-domain image-to-image translation,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 404–419.

W. Chu, Y. Tai, C. Wang, J. Li, F. Huang, and R. Ji, “SSCGAN: Facial attribute editing via style skip connections,” in Proc. Eur. Conf. Comput. Vis., 2020.

X. Chen et al., “CooGAN: A memory-efficient framework for high-resolution facial attribute editing,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 670–686.

B. Bhattarai and T.-K. Kim, “Inducing optimal attribute representations for conditional GANs,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 60–85.

G. Doria, S. Vicente, N. D. Campbell, and L. J. Simpson, “The GAN that warped: Semantic attribute editing with unpaired data,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5356–5365.

Y. Gao et al., “High-fidelity and arbitrary face editing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 16115–16124.

S. Gu, J. Bao, D. Chen, and F. Wen, “GIQA: Generated image quality assessment,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 369–385.

X. Yao, G. Puy, A. Newson, Y. Gousseau, and P. Hellier, “High resolution face age editing,” in Proc. Int. Conf. Pattern Recognit., 2021, pp. 8624–8631.

M.-Y. Liu et al., “Few-shot unsupervised image-to-image translation,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 10551–10560.

M. Afifi, M. A. Brubaker, and M. S. Brown, “HistoGAN: Controlling colors of GAN-generated and real images via color histograms,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 7941–7950.

T. Portenier, Q. Hu, A. Szabó, S. A. Bigdeli, P. Favaro, and M. Zwicker, “FaceShop: Deep sketch-based face image editing,” ACM Trans. Graph., vol. 37, no. 4, p. 174:1–174:12, 2018.

Y. Jo and J. Park, “SC-FEGAN: Face editing generative adversarial network with user’s sketch and color,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 1745–1753.

Z. Tan et al., “MichGAN: Multi-input-conditioned hair image generation for portrait editing,” ACM Trans. Graph., vol. 39, no. 4, pp. 95:1–95:13, 2020.

H.-P. Wang, N. Yu, and M. Fritz, “Hijack-GAN: Unintended-use of pretrained, black-box GANs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 7872–7881.

Y. Jiang, Z. Huang, X. Pan, C. C. Loy, and Z. Liu, “Talk-to-edit: Fine-grained facial editing via dialog,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 13799–13808.

P. Zhang, O. O. Koyejo, and A. Schwing, “Enjoy your editing: Controllable GANs for image editing via latent space navigation,” in Proc. Int. Conf. Learn. Representations, 2021.

Y. Xu, Y. Shen, J. Zhu, C. Yang, and B. Zhou, “Generative hierarchical features from synthesizing images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2021, pp. 4432–4442.

J. Han, Z. Chen, Y. Xia, S. Liu, T. Qin, and J. Luo, “Exploring explicit domain supervision for latent space disentanglement in unpaired image-to-image translation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 4, pp. 1254–1266, Apr. 2021.

A. Tewari et al., “PIE: Portrait image embedding for semantic control,” ACM Trans. Graph., vol. 39, no. 6, pp. 1–14, 2020.

H. Ling, K. Kreis, D. Li, S. W. Kim, A. Torralba, and S. Fidler, “EditGAN: High-precision semantic image editing,” in Proc. Int. Conf. Neural Inf. Process. Syst., 2021, pp. 16331–16345.
Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
Qi Li (Member, IEEE) received the BE degree from the China University of Petroleum, in 2011, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2016. He is an associate professor with the Center for Research on Intelligent Perception and Computing (CRI PAC), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), Institute of Automation, Chinese Academy of Sciences (CASIA). His research interests include face recognition, computer vision, and machine learning.

Qiayao Deng received the BE degree in automation from the Beijing University of Chemical Technology, in 2017, and the PhD degree in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2022. She is a lecturer with People's Public Security University of China. Her research interests include biometrics, pattern recognition, computer vision, and machine learning.

Zhenan Sun (Senior Member, IEEE) received the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2006. He is a professor with the Center for Research on Intelligent Perception and Computing (CRI PAC), State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS), CASIA. His current research interests include biometrics, pattern recognition, and computer vision. He is a fellow of the IAPR, and an associate editor of IEEE Transactions on Biometrics, Behavior, and Identity Science.

Ming-Hsuan Yang (Fellow, IEEE) is a professor with the University of California at Merced and Yonsei University. He received the Longuet-Higgins Prize, in 2023, NSF CAREER Award, in 2012, and Google Faculty Award, in 2009. He is a fellow of the ACM.