Unpaired Multi-domain Attribute Translation of 3D Facial Shapes with a Square and Symmetric Geometric Map

Zhenfeng Fan[1,2], Zhiheng Zhang[1,2], Shuang Yang[1,2], Chongyang Zhong[1,2], Min Cao[3], and Shihong Xia[1,2*]

[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Soochow University
{fanzhenfeng,zhangzhiheng20g,yangshuang21b,zhongchongyang,xsh}@ict.ac.cn; mcao@suda.edu.cn

Abstract

While impressive progress has recently been made in image-oriented facial attribute translation, shape-oriented 3D facial attribute translation remains an unsolved issue. This is primarily limited by the lack of 3D generative models and ineffective usage of 3D facial data. We propose a learning framework for 3D facial attribute translation to relieve these limitations. Firstly, we customize a novel geometric map for 3D shape representation and embed it in an end-to-end generative adversarial network. The geometric map represents 3D shapes symmetrically on a square image grid, while preserving the neighboring relationship of 3D vertices in a local least-square sense. This enables effective learning for the latent representation of data with different attributes. Secondly, we employ a unified and unpaired learning framework for multi-domain attribute translation. It not only makes effective usage of data correlation from multiple domains, but also mitigates the constraint for hardly accessible paired data. Finally, we propose a hierarchical architecture for the discriminator to guarantee robust results against both global and local artifacts. We conduct extensive experiments to demonstrate the advantage of the proposed framework over the state-of-the-art in generating high-fidelity facial shapes. Given an input 3D facial shape, the proposed framework is able to synthesize novel shapes of different attributes, which covers some downstream applications, such as expression transfer, gender translation, and aging. Code at https://github.com/NaughtyZZ/3D_facial_shape_attribute_translation_ssgmap.

*The corresponding author

1. Introduction

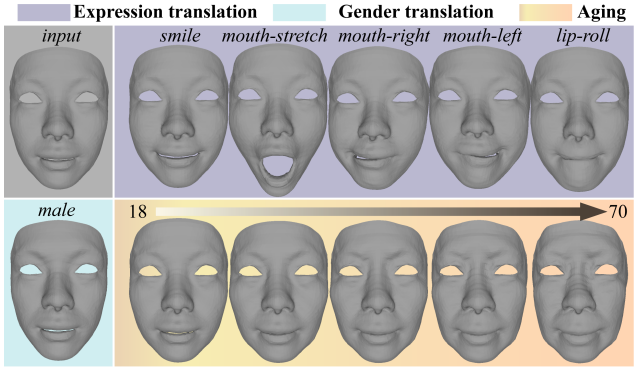

Figure 1. Translation of 3D shape attributes to multiple domains that correspond to expressions, genders, and ages.

The advancement of generative adversarial networks (GANs) has recently activated a lot of studies on face-related tasks, such as face synthesis [34, 7], facial image super-resolution [38, 32], and facial attribute translation [13, 1]. The rationale of these studies is to use an auxiliary discriminator network to regularize the output face with the learned distribution of facial data towards a specific category. The task of facial attribute translation is referred to as changing the particular aspect of a facial image, *e.g.* changing the expression [19] or the age [61] of a face, resulting in desired appearances. This task has many downstream applications in the media and film industry.

In the past, most methods in the field of facial attribute translation dealt with 2D facial images [11, 51, 55]. This is partly because 2D facial images are easily accessible and partly because deep learning based methods commonly work on image data with a regular grid. Nowadays, re-

searchers are paying more attention to 3D faces [18, 57, 50, 17, 21, 67, 49] for wider applications and more realistic rendering, due to advances in 3D imaging sensors and 3D applications. Editing 3D facial shape has attracted much interest in both the computer graphics and computer vision communities. Translation of 3D facial shapes provides geometric flexibility in addition to textures, thus resulting in vividness for facial rigging and animation [23, 40, 53, 25]. The 3D shape is also considered as a vital commodity of face to overcome pose and illumination challenges in existing face recognition literature [42, 65, 43].

In this work, we study the problem of 3D facial shape attribute translation. An example is shown in Figure 1. We denote *attributes* as inherent features of a 3D face, such as **expression**, **age**, and **gender** that correlate to shape variations. We also refer to some GAN-based domain adaptation works [13, 31], and define "*domain*" as a set of data with certain attribute values, *e.g.* 30-year-old males with neutral expression. We cast 3D facial attribute translation as a domain adaptation problem, which is naturally linked with GAN by its data-driven nature.

Applying state-of-the-art deep GANs on 3D geometry data is challenging, and the difficulties are mainly two-fold. Firstly, unlike facial images on a square Euclidean grid, 3D geometric data, in the case of facial surfaces, are on a Riemannian manifold. This hinders the application of state-of-the-art deep convolutional neural networks (CNNs) on 3D facial attribute translation. We refer to this as a problem of *network compatibility*. Secondly, unlike facial image data that are abundant for image translation tasks, there is a shortage of 3D facial data. This is limited by the popularity of 3D scanning devices. Moreover, most deep CNN based methods rely heavily on paired and labeled data. We refer to this as a problem of *data scarcity*.

For *network compatibility*, we design a geometric map that encodes 3D coordinates onto regular image grids. The adjacency information for 3D vertices is preserved in a *local least-square* manner while being constrained by *symmetric* property. This enables us to leverage symmetry, an important character of face in the learning process. In addition, the learning networks are *end-to-end* with a differentiable 3D-to-2D forward geometric mapping layer and a 2D-to-3D backward grid sampling layer.

For *data scarcity*, we employ a *unified* and *unpaired* GAN for multi-domain attribute translation. Firstly, we assume that the latent encodings of different domains should be cross-correlated to each other. Rather than learning the translation between every two domains separately, we learn a single generator for all domains, *i.e.* for expression, age, and gender translation tasks together. Secondly, we employ an unpaired framework, assuming that exactly paired data, *i.e.* different ages of a person are difficult to collect and different genders for the same identity are almost impossible in the real world. This mitigates the exact constraint for hardly accessible paired data. We also conduct data augmentation in training by adding random perturbations of scales and rotations of 3D facial shapes.

In summary, the main contributions of this paper are:

- We first propose a general and unified framework for multi-domain 3D facial attribute translation, which covers some shape-oriented applications including expression transfer, aging, and gender translation.
- We construct a novel geometric map for 3D face representation on a canonical 2D grid. The geometric map leverages symmetry of face and maintains the adjacency of 3D vertices in a local least-square manner.
- We make unpaired training of 3D facial shape data available on a geometric map with a hierarchical GAN architecture to suppress both global and local artifacts.

2. Related Work

The closely related fields to this work include GAN-based 2D image translation and 3D face manipulation with UV maps. We now briefly discuss the most related works in each field, respectively.

2.1. GAN-based 2D Image Translation

GAN is very popular for generating novel and realistic images [32, 44, 64, 66] because of its capability for modeling the distribution of a large amount of data. Many variants of GANs with different characteristics are proposed after the seminal work of Vanilla GAN [24].

Conditional GANs (CGANs) [46, 48] are originally proposed to generate samples conditioned on a specific class. CGANs commonly include class labels in both the generator and discriminator that are relevant to a specified task. For example, pix2pix [32] learns image-to-image translation with a CGAN architecture in a supervisory manner with paired data; Attgan [29] is able to change the specified attribute of a facial image.

CycleGAN [68] and DisCoGAN [36] release the paired data assumption by incorporating a cycle consistency loss that preserves the key attribute shared by input and output images. However, these GANs are limited to attribute translation tasks between two domains. Thus they cannot effectively make use of data that cover multiple domains. To alleviates the deterministic mapping problem, Huang *et al.* propose an MUNIT framework [30] by incorporating a style code to generate versatile images.

StarGANs [13, 14] can learn multi-domain attribute translation with the help of domain classification loss in addition to cycle consistency loss and adversarial loss. Generally, they can effectively learn both the global features shared by data in all domains and the local features hold

by data in a specific domain. U-GAT-IT [35] and Attentiongan [56] further employ the attention mechanism to generate high-quality foreground against background.

In this work, we borrow some key architectures in StarGANs [13, 14] for 2D images and design a geometric map elaborately for 3D facial data. We further incorporate them in an end-to-end adversarial learning framework with a novel hierarchical discriminator. Therefore, the proposed framework supports multi-domain 3D facial attribute translation in an unpaired manner with a single generator.

2.2. UV Maps for 3D Face Manipulation

In a departure from 2D images with pixels on a regular grid, raw 3D geometric data are commonly organized as irregular points. A common way to represent 3D facial data is to flatten the 3D surface onto a 2D UV plane, on which the locations of 3D vertices are encoded.

Blanz and Vetter [9] project the facial surface onto a cylindrical UV map for shape registration in their seminal work for 3D Morphable Model (3DMM). The 3D location of each vertex is encoded as height and distance to the cylindrical axis. The 3DMM is further used for an attribute translation task to manipulate weight, gender, and expressions [3].

Bagautdinov *et al.* [6] conduct non-rigid registration of 3D facial scans and deform a template face (as the average of all registered results) to a 2D plane. They further define a 3-channel image (UV map) surrounding the deformed template. This ensures topological neighbors on 3D are also topological neighbors on the image. This strategy to represent 3D facial data is used by many recent works [2, 22, 45, 47] for 3D faces. Generally, the locations of 3D vertices can be converted into an image-like tensor to apply the 2D convolutions [22]. An advantage over some other works with voxel/point/graph convolution [33, 52, 12, 41] is the high-frequency details can be preserved better.

In this work, we construct a novel UV map that supports forward and backward differentiable operations to be embedded in an end-to-end learning network. We call our UV map a geometric map because it has the following novel geometric properties: 1) It is square; 2) It is symmetric with respect to a central axis; 3) the mapping between 3D vertices and their 2D correspondences is in a local least-square manner to preserve the 3D adjacency information as rigidly as possible. These properties enable us to learn a 3D attribute translation task effectively.

3. Method

In this section, we describe the proposed geometric map, the network architecture, and the detailed loss functions in training, respectively, which constitute the basic elements of the proposed framework (see Figure 2).

3.1. Geometric Map Construction

We consider triangle facial meshes as the input and output of the proposed method with the same number of vertices and the same mesh topology. Generalization to other formats of data, *e.g.* point cloud, is applicable by rigidly alignment [8, 63] and non-rigid registration [4, 10] to a common template mesh in our implementation. Since all faces share the same topology, we denote each face by $\mathcal{V} \in \mathbb{R}^{3\times n}$ for simplicity. In order to make 2D deep CNN architectures compatible with 3D facial shapes, we construct a geometric map that bridges the gap between a 3D surface and a 2D image grid. The steps are:

I. Average over some registered facial meshes to acquire a noiseless 3D template mesh $\mathcal{V}^s$;

II. Initialize the geometric map by harmonic parametrization [16, 26] of $\mathcal{V}^s$ to $\mathcal{V}^t$;

III. Deform $\mathcal{V}^t$ to a **square and symmetric** geometric map guided by the local structure of $\mathcal{V}^s$ and some rearranged key vertices, as follows (also refer to Fig. 3).

First, we select some key vertices on $\mathcal{V}^t$ that are landmarks, edges, and central axis on the face. We adjust the central vertices to form the central axis. The locations of the landmarks and edges are rearranged to be square and symmetric with respect to the central axis.

Then, we denote the 1-ring neighbors of each vertices $v_i^s \in \mathcal{V}^s$ as $\mathcal{N}^1(v_i)$. The correspondence of each vertex in $\mathcal{V}_s$ to $\mathcal{V}_t$ is determined by the subscript i. We suppose there exists a rigid transformation $\{R_i, T_i\}$ that aligns v_i^s to v_i^t, as

$$v_i^t \leftarrow R_i v_i^s + T_i (i \in \mathcal{V}^s), \tag{1}$$

where $\{R_i, T_i\}$ can be estimated by a least-square alignment problem of the surrounding 1-ring neighbors, as

$$\{R_i, T_i\} = \underset{R_i \in SO(3), T_i \in \mathbb{R}^3}{\arg\min} \sum_{v_j^s \in \mathcal{N}^1(v_i)} \left\| R_i v_j^s + T_i - v_j^t \right\|_2^2. \tag{2}$$

Here $SO(3)$ denotes the space of all Givens matrices.

Next, the preliminary offset for each vertex is regularized by local smoothness. We denote $p_i^t = R_i v_i^s + T_i$ as the unregularized offset and formulate the problem as

$$\begin{aligned} \{o_i | i \in \mathcal{V}^t\} = \underset{\{o_i | i \in \mathcal{V}^t\}}{\arg\min} \{ & \sum_{i \in \mathcal{V}^t} \left\| p_i^t - (v_i^t + o_i) \right\|_2^2 \\ & + \sum_{i \in \mathcal{V}^t} \sum_{j \in \mathcal{N}^1(v_i)} \| o_i - o_j \|_2^2 \}. \end{aligned} \tag{3}$$

Solving Eq. 3 requires taking the partial derivative with respect to each offset $o_i (i \in \mathcal{V}^t)$ and leads to a linear system

$$[\mathbf{A}_{ij}]_{n\times n} \cdot [\mathbf{O}_{ij}]_{n\times 3} = [\mathbf{B}_{ij}]_{n\times 3}, \tag{4}$$

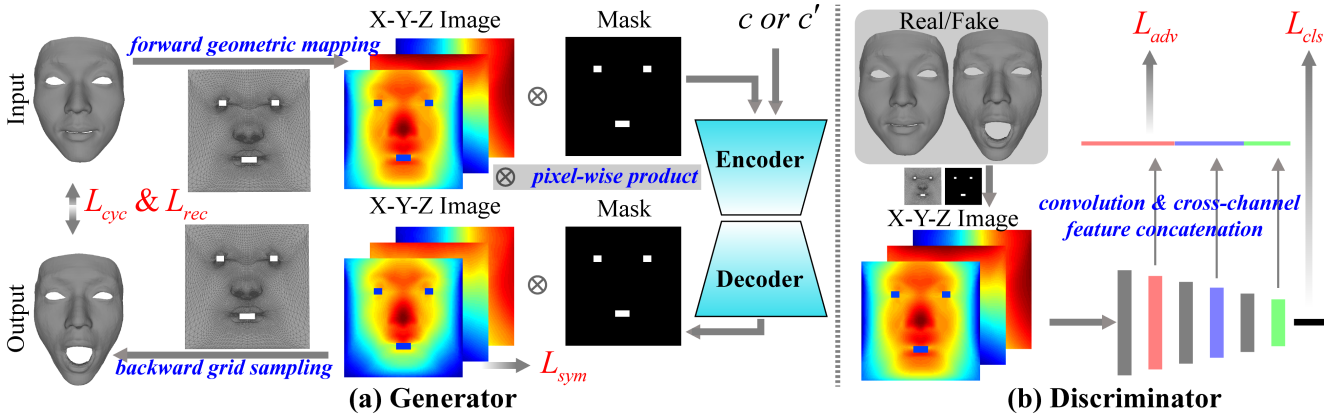

Figure 2. The overall network architecture with the proposed geometric map. The loss functions for training are marked in red typeface.

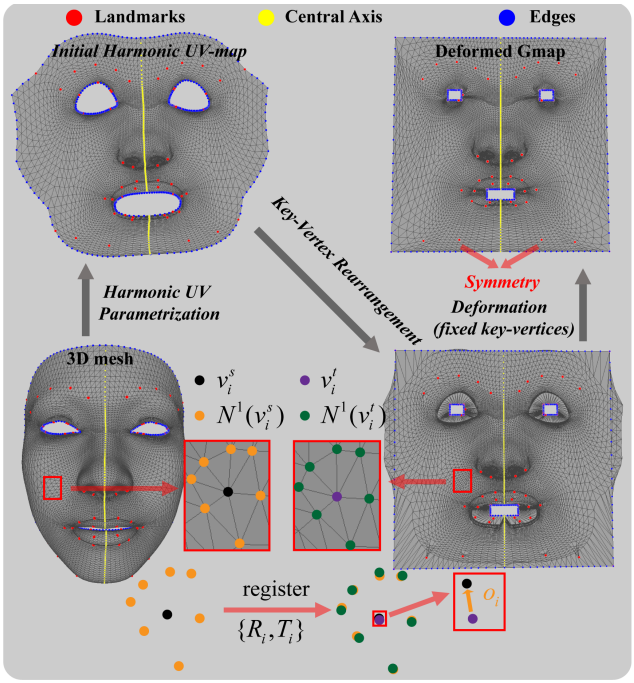

Figure 3. The process to deform an initial harmonic UV-map to a square and symmetric geometric map (**Gmap**).

where

$$\mathbf{A}_{ij} = \begin{cases} 1 + 2N_{v_i} & \text{if} \quad i = j \\ -2 & \text{if} \quad i \neq j \text{ and } j \in \mathcal{N}^1(v_i) \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

$$[\mathbf{O}_{ij}]_{n\times 3} = [o_1, ..., o_n]^T, \tag{6}$$

and

$$[\mathbf{B}_{ij}]_{N\times 3} = [p_1^t - v_1^t, ..., p_n^t - v_n^t]^T. \tag{7}$$

N_{v_i} is the number of vertices in $\mathcal{N}^1(v_i)$ and the superscript T denotes matrix transpose.

Since some key vertices on the **landmarks, edges, and central axis** are fixed as constraints for solving Eq. 3, we exclude the corresponding columns in $\mathbf{A}$ and rows in $\mathbf{O}$. Therefore, $\mathbf{A}$ is a *rank-deficient* matrix. Let n_f be the number of fixed vertices ($\mathcal{V}^f \subset \mathcal{V}^t$), then Eq. 4 is degraded to

$$[\mathbf{A}_{ij}]_{n\times(n-n_f)} \cdot [\mathbf{O}_{ij}]_{(n-n_f)\times 3} = [\mathbf{B}_{ij}]_{(n-n_f)\times 3}, \tag{8}$$

which is an *over-determined* linear system. Its *least-square* solution is given by the *Moore-Penrose inverse* as

$$\mathbf{O} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{B}. \tag{9}$$

By this way, we are able to fix some key vertices while updating other vertices. The vertex on the target mesh is added by each offset in $\mathbf{O}$ after solving Eq. 8, as

$$v_i^t = v_i^t + o_i (i \in \mathcal{V}^t / \mathcal{V}^f). \tag{10}$$

Finally, the steps from Eq. 1 to Eq. 10 are iterated until the ensembled offset is smaller than a certain threshold.

The above process is in fact a variant of a locally rigid registration process in [20]. It conducts as rigid as possible alignments from the local cells (1-rings) on the original 3D mesh to the deformed geometric map, while being constrained by some rearranged key vertices. It also demonstrates that the proposed geometric map preserves the adjacency relationship[1] of all vertices in a *least-square* manner. This makes the mapping from 3D mesh to 2D geometric map to be one-to-one for each vertex, avoiding triangle flipping which is correlated to interpolation errors. Furthermore, as described in [26, 54], sampling the locations in a geometric image is prone to large interpolation errors unless the border vertices are preassigned to distinct pixels. In this work, there is seldom triangle flips even on the border regions, thus avoiding most interpolation errors.

[1]This brings about an advantage over the existing UV maps [6, 2, 60]. The vertices on the upper and lower mouth area are separated, avoiding interferences from the local convolutional kernels in common CNNs.

In addition, the advantages for the geometric map to be *square and symmetric* are: 1) making full use of the pixels (2D images also cover a square area); 2) easy to exclude the blank pixels inside the eyes and mouth; 3) easy to leverage a symmetric loss in training by image flip operation.

3.2. Network Architecture

The network architecture[2] employs a main generator network and an auxiliary discriminator network, as in Figure 2. The mapping between 3D vertices and 2D image grid is computed by *forward barycentric interpolation* (as in [2]) and *backward bilinear grid sampling* with the aforementioned geometric map[3]. A mask for the valid pixels is customized accordingly. In a departure from some existing works, the forward and backward mappings are incorporated into our network in an end-to-end manner. During training, the generator network transfers the 3D facial shape to a certain domain, while the discriminator network enforces indistinguishable generation (probability distribution) to the given domain.

Generator. We use an encoder-decoder architecture for the generator network. The feature expansion of the bottleneck is only downsampled by a factor of 4 to avoid the mixing of spatial content (thus friendly to high-frequency information) of the input geometric map. We also include 6 residual blocks [28] in the bottleneck. The skip connection of the residual block is able to learn the latent representations of data in multiple domains efficiently. Moreover, we embed both the forward geometric mapping layer and backward grid sampling layer as the input and output layers of the network, respectively. The end-to-end learning setting can compensate for the sampling errors between a 3D face and its representation on the geometric map.

Discriminator. Previous studies discriminate the feature in different dimensions, *e.g.* Vanilla GAN [24] for the global image, Pixel-GAN [32] for each pixel, and Patch-GAN [39] for a few intersected receptive fields on an image. In a departure from that, we employ a strategy that we call Pyramid-GAN to discriminate both global and local patterns of the input shape. The discriminator downsamples the feature expansion by 2 in a cascade manner until the height and width are 2×2. Specifically, we conduct bifurcated convolutional operations on every other feature layer, then flatten and concatenate each output, and finally fed them to the adversarial loss. This strategy leads to both globally and locally realistic generations of 3D faces.

3.3. Loss Function

We employ several loss functions (also refer to Figure 2 for the exact locations) in the training process. The symbols of some variables are listed below for brevity.

- x denotes the input shape of the generator fed into the forward geometric mapping layer.
- y denotes the output geometric map of the generator fed into the backward grid sampling layer.
- c and c' are the target and source domain labels of x, respectively, which are composed of discrete expression and gender labels and continuous age labels.
- G denotes the generator network; D_{src} and D_{cls} denotes the real/fake and domain classification branch of the discriminator network, respectively.

The Adversarial Loss forces the distribution of the generated 3D shapes to approach the shapes of real faces, as

$$L_{adv} = E_x[\log D_{src}(x)] + E_{x,c}[\log(1 - D_{src}((G(x,c))], \tag{11}$$

given the generator taking as input 3D shape x conditioned on domain label c. In practice, we adopt the Wasserstein GAN [5] with gradient penalty to stabilize the training process as

$$\begin{aligned} L_{adv} =& E_x[\log D_{src}(x)] - E_{x,c}[D_{src}((G(x,c) \\ & - \lambda_{gp} E_{\hat{x}}[(\|\nabla_{\hat{x}} D_{src}(\hat{x})\|_2 - \alpha)^2]], \end{aligned} \tag{12}$$

where $\hat{x}$ is uniformly sampled between a pair of real and generated shapes. And we set α to 0.01.

The Classification loss enables effective domain translation from c' to c. The objective is decomposed into

$$L^r_{cls} = E_{x,c'}[-\log D_{cls}(c'|x)] \tag{13}$$

for real shapes to optimize the discriminator, and

$$L^f_{cls} = E_{x,c}[-\log D_{cls}(c|G(x,c))] \tag{14}$$

for fake shapes to optimize the generator. In practice, we employ cross-entropy loss for discrete expression or gender parts and mean-square loss for the continuous age part of the domain label.

The Cycle loss employs a cycle consistency loss [68] to preserve the domain-unrelated part while changing only the domain-related part in the input image, as

$$L_{cyc} = E_{x,c,c'} \left\| x - G(G(x,c),c') \right\|_1 . \tag{15}$$

The Reconstruction loss makes the generator stable if the target domain label remains unchanged with respect to the source domain, as

$$L_{rec} = E_{x,c'} \left\| x - G(x,c') \right\|_1 . \tag{16}$$

The Symmetry loss enforces the output facial shapes to be symmetric with respect to the central axis, as

$$L_{sym} = E_y \left\| y - F(y) \right\|_1 , \tag{17}$$

[2]The detailed architecture is described in the supplementary material.

[3]The detailed exposition for the forward and backward mappings is described in the supplementary material.

where F is a flip operator that flips horizontally for y and then reverses the sign for the first channel. **We exclude the symmetric loss for asymmetric expressions in training.**

The full objective functions to optimize G and D are weighted combinations of the above loss terms as

$$L_D = -L_{adv} + \lambda_{cls} L^r_{cls} \tag{18}$$

and

$$L_G = L_{adv} + \lambda_{cls} L^f_{cls} + \lambda_{cyc} L_{cyc} + \lambda_{rec} L_{rec} + \lambda_{sym} L_{sym}, \tag{19}$$

respectively.

4. Experiments

In this section, we carry out experiments on a public dataset [62] to demonstrate the effectiveness of the proposed method[4]. We also show that the trained model is generalizable to other datasets [15, 59] combined with proper registrations.

4.1. Dataset & Labels

The FaceScape dataset [62] provides a set of topological uniform (registered) samples with the NICP [4] method. These samples cover 847 identities, 20 expressions, age ranges from 16 to 70, and 2 different genders. We select $14,486$ out of all $18,760$ samples to exclude some noisy registered results. Among the selected samples, we further choose $10,005$ samples for training, 32 samples as a mini-batch for validation using the *leave-one-out principle* [27], and the rest samples for testing. The identities of the training and testing samples are disjoint.

One advantage of the proposed method is that it learns multi-domain translation with a unified GAN. To this end, we customize a domain label vector $c = \{c_i | i = 1, 2, ...23\}$ of length 23. The first 20 dimensions and the following $21-22$ dimensions are **one-hot binary** expression and gender label, respectively. And the last dimension is a normalized **one-hot number** ranged by $[-1, 1]$ for the ages.

4.2. Training Details & Hyper-parameter Selection

We implement all the networks with the Pytorch platform. The generator and discriminator are alternated *one-by-one* in the training process. Adam [37] optimizer is used. The total iterations (mini-batches) are set to $800,000$. The first $400,000$ iterations adopt a fixed learning rate of $1e-4$. In the last $400,000$ iterations, the learning rate decreases to $1e-6$ linearly for every $2,000$ iterations. It takes about 50 hours on a single GPU (specified as NVIDIA RTX 3090) to train the proposed networks. In the training process, we augment the data by random scaling and rotation by a range of $[0.9, 1.1]$ and $[-10', 10']$ for 3 Euler angles, respectively.

The hyper-parameter settings for the weights of different loss functions in Eq. 12 and Eq. 19 follow a *trial-and-error* principle guided by the visualized results of the validation set. We also follow some general settings in existing works [5, 13] and adjust the parameters until the loss curves keep oscillating, as a common way for tuning GAN. In our experiment, we set $\lambda^C_{cls} = 0.02$, $\lambda^M_{cls} = 0.05$, $\lambda_{cyc} = 2$, $\lambda_{rec} = 0.1$, $\lambda_{sym} = 0.5$, and $\lambda_{gp} = 0.2$, where the superscripts C and M on λ_{cls} denote binary cross-entropy loss for classification (for expression and gender labels) and mean-square error loss for regression (for age label), respectively.

[4]We use the FaceScape dataset to train our model because it is the only one so far that provides registered 3D shapes with various attributes.

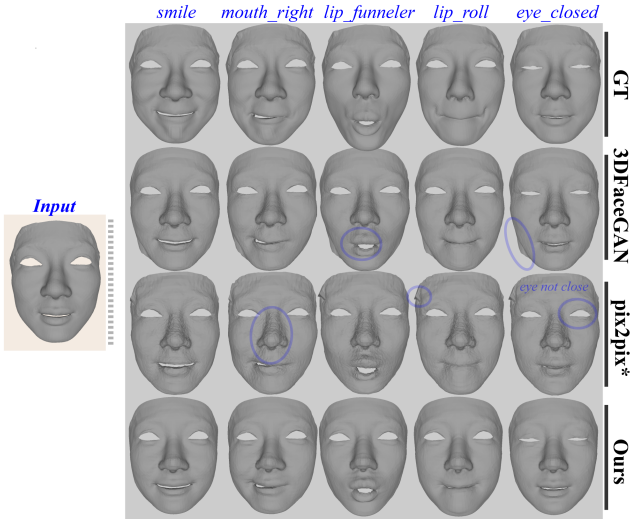

Figure 4. Qualitative comparison for an exemplar input face. "GT" denotes the ground truth. Some artifacts and failure cases by the two baseline methods are marked with circles.

4.3. Comparison for Expression Translation

Previous works on deep learning based attribute translation mostly focus on 2D images [13] or 3D texture [60]. There is seldom work on 3D facial shape attribute translation. However, there are some traceable works on 3D face generation with expressions. Therefore, we implement two baseline models which are close to ours to the best of our knowledge for comparisons. Since the UV map in [47] cannot be reproduced exactly, we use our proposed geometric map instead.

- **3DFaceGAN** [47] is the first GAN tailored towards modeling the distribution of 3D facial surfaces. It can be used for 3D facial expression translation, which employs Multivariate-Gaussian decomposition and supervision to handle expressions on a UV map.
- **Pix2pix** [32] is a widely used GAN for image trans-

lation applications. We blend our proposed geometric map with the official implementation [32] for pix2pix. We also constructed paired 3D shape data for neutral and other expressions for training.

Method	MSE-V (mm)	MSE-N (degree)
3DFaceGAN [47]	1.05	0.99
Pix2pix* [32]	2.33	3.47
Our work	**0.84**	**0.51**

Table 1. Comparisons of MSE on vertices and normal. The mark * denotes a combination with the proposed geometric map.

We select neutral expressions for all subjects and transfer it to other expressions in the test set. After that, we compare the per-vertex and corresponding normal mean square error (**MSE-V** and **MSE-N**) between the generated results and the ground-truth ones after rigid registration of each pair of them by iterative closest point (**ICP**) [8] method. Table 1 and Figure 4 show the quantitative and qualitative comparisons. We can see that our proposed method performs over the pix2pix method by a large margin even without the paired setting. This is mainly due to the effective usage of data from multiple domains. In contrast, the pix2pix model only makes use of limited data pairs from two domains. Although the improvement of our results on MSE over 3DFaceGAN is not very salient, our qualitative results are visually better, especially on the edges of the mesh. We owe our success to the multi-domain translation framework and proper loss functions. Note that a result that is different from the ground truth is reasonable, since our method is supervised only by expression labels rather than pixel-wise errors.

4.4. Multi-domain Translation

In addition to expression transfer, our proposed framework is able to manipulate age and gender continuously[5], which is a notable advantage over the existing works. Specifically, we only train a single network for expression, age, and gender translation tasks. Furthermore, the proposed method also supports input in various domains that are not limited to neutral such that it can broaden the applications to produce user-defined attributes (see Figure 5).

4.5. Clustering Results for Feature Representation

In this experiment, we project the output feature vectors at the classification head of the discriminator network onto a latent 2D plane to view the clustering effect of data. We employ a t-distributed stochastic neighbor embedding (t-SNE) [58] method for latent space visualization. Figure 6 shows the results marked with expression, age, and gender labels for some randomly generated samples from the testing samples in the FaceScape dataset. We can see that different classes are well separated and show noteworthy clustering effects, which implies that the trained model learns the representation of data of different domains effectively. Note that the three sub-figures for different attributes share the same representations.

[5]Please refer to Figure 1 and more qualitative results on the supplementary material.

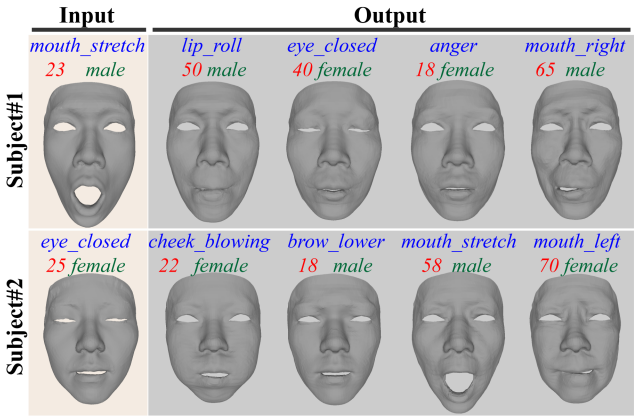

Figure 5. Two qualitative examples for outputs with various attributes which are mostly different from the inputs.

4.6. Ablation Studies

We now conduct experiments to learn the effects of different components of the proposed framework. The generator architectures is borrowed from the starGANs [13, 14]. Some other components that are new to the existing works include:

- The square and symmetric geometric map. We replace it with the plain harmonic UV-map instead for the ablation study. This considers that the plain UV map is used in many existing works [60, 47].
- The Pyramid-GAN architecture. We replace the Pyramid-GAN architecture with Patch-GAN [39, 13] only in the second last feature layer for ablation study. The Pyramid-GAN is in fact a generalization of Patch-GAN to diverse feature expansions.
- The symmetric loss function. We neglect this loss function for the ablation study.
- The reconstruction loss function. We neglect this loss function for the ablation study.

We train the networks with the above settings, respectively. Table 2 shows the quantitative results in terms of the aforementioned MSE scores. We can see that each element contributes to some gains, among which the geometric map and the identity loss take effects most. Some qualitative results are also selected for a better understanding of the

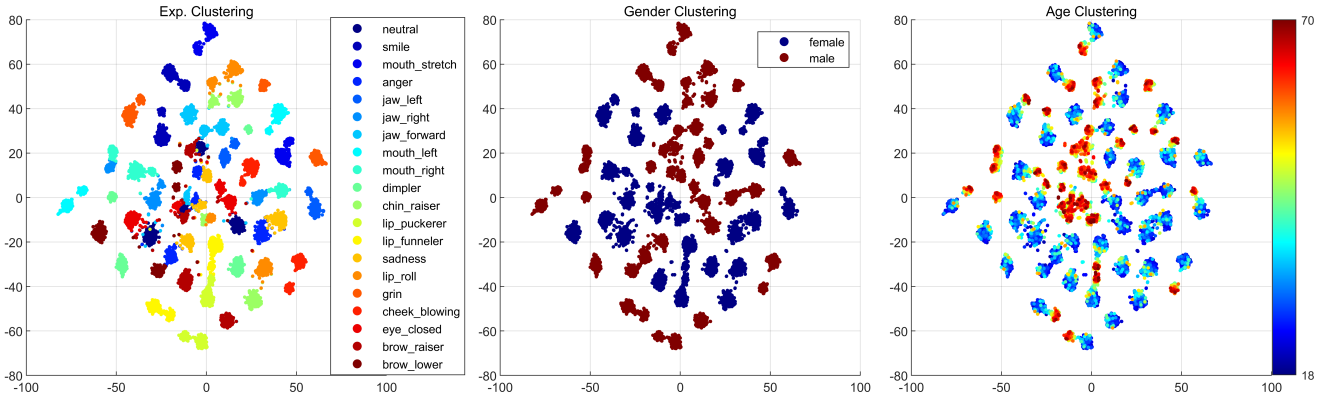

Figure 6. Clustering of generated data with the t-SNE [58] method for different expressions, genders, and ages, respectively.

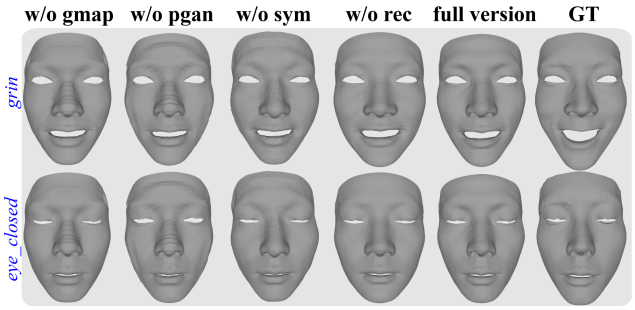

Figure 7. Qualitative results of 2 generated expressions for an exemplar face. The ground truths are also included for reference.

Method	MSE-V (mm)	MSE-N (degree)
W/O geometric map	1.31	1.17
W/O pyramid GAN	1.06	1.13
W/O symmetric loss	0.91	0.78
W/O reconstruction loss	2.79	1.98
Full version	**0.84**	**0.51**

Table 2. Comparisons on MSE for the ablation study.

underlying mechanism of each proposed component, as in Figure 7. We observe that: 1) the geometric map contributes to realistic details since it handles a fundamental one-to-one 3D-to-2D mapping problem; 2) the Pyramid-GAN architecture reduces some artifacts over Patch-GAN; 3) the symmetric loss utilizes some prior knowledge of face for more stable and noiseless results; 4) the reconstruction loss resists identity drift. It is worth noting that the identity drift is not common in image-oriented tasks, however, it is a problem in the shape-oriented task in this work. This is because the pixel-wise values rather than 2D patterns are correlated to 3D shapes directly.

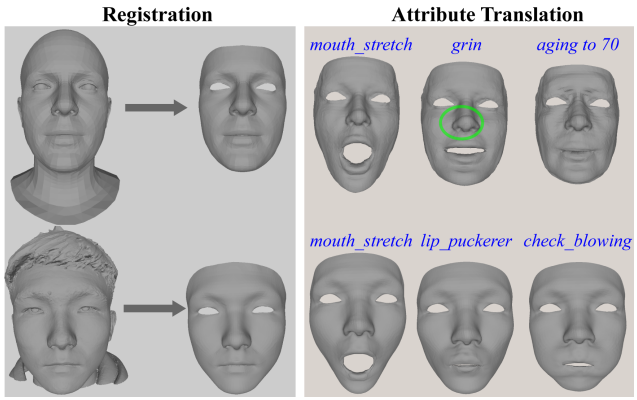

Figure 8. Generalization to two exemplar faces from VOCA [15] and FaceVerse [59] datasets, respectively.

4.7. Generalization Test & Limitations

The deep learning methods on image data often suffer from generalization problems. However, in this work, we find that the trained model on FaceScape is well generalizable to other datasets if the target face is registered to a common template properly. We register the template to some samples from two other datasets (VOCA [15] and FaceVerse [59]), and then feed the registered samples into the trained model. Figure 8 shows two examples for attribute translation. Overall, both of the two samples are translated to the target attributes well. The reason should be that facial shape data are much simpler than texture data, since 3D shapes are immune to illumination and pose changes. However, some unexpected artifacts appear on the nose region (marked with green circles) in the top example. This is due to the intrinsic bias of the training set, *i.e.*, all training samples are east Asian while the testing example is Caucasian. In the future, we will try to train a model with richer data that cover various ethnicities for better generalization.

5. Conclusion

In this paper, we propose an unpaired end-to-end adversarial learning framework for multi-domain 3D facial shape attribute translation. Given an input 3D facial shape, the proposed framework is capable of synthesizing realistic 3D facial shapes with various expressions, genders, and ages with a unified generator network. The key element of our proposed framework is the canonical representation of 3D faces by a square and symmetric geometric map, enabling effective learning on facial surfaces. Others include a Pyramid-GAN architecture and task-related loss functions, enabling unified and unpaired training with 3D data on the geometric map robustly. Extensive experiments demonstrate the effectiveness of the proposed method for the translation of various facial attributes. We hope this work will be helpful for future research and applications.

Acknowledgement. We would like to thank all the anonymous reviewers for their insightful comments. This work is supported in part by the National Key Research and Development Program of China (No. 2022YFF0902302), the National Science Foundation of China (No. 62106250 and No. 62002252), and China Postdoctoral Science Foundation (No. 2021M703272).

References

[1] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. *ACM Transactions on Graphics*, 40(3):1–21, 2021. 1

[2] Victoria Fernández Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, and Edmond Boyer. A decoupled 3d facial shape model by adversarial training. In *IEEE International Conference on Computer Vision*, pages 9419–9428, 2019. 3, 4, 5

[3] Brian Amberg, Pascal Paysan, and Thomas Vetter. Weight, sex, and facial expressions: On the manipulation of attributes in generative 3d face models. In *International Symposium on Advances in Visual Computing*, pages 875–885, 2009. 3

[4] Brian Amberg, Sami Romdhani, and Thomas Vetter. Optimal step nonrigid icp algorithms for surface registration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007. 3, 6

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017. 5, 6

[6] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 3, 4

[7] Linchao Bao, Xiangkai Lin, Yajing Chen, Haoxian Zhang, Sheng Wang, Xuefei Zhe, Di Kang, Haozhi Huang, Xinwei Jiang, Jue Wang, et al. High-fidelity 3d digital human head creation from rgb-d selfies. *ACM Transactions on Graphics*, 41(1):1–21, 2021. 1

[8] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. 3, 7

[9] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Annual conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 3

[10] James Booth, Anastasios Roussos, Allan Ponniah, David Dunaway, and Stefanos Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126(2):233–254, 2018. 3

[11] Jia-Ren Chang, Yong-Sheng Chen, and Wei-Chen Chiu. Learning facial representations from the cycle-consistency of face. In *IEEE International Conference on Computer Vision*, pages 9680–9689, 2021. 1

[12] Shiyang Cheng, Michael Bronstein, Yuxiang Zhou, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. Meshgan: Non-linear 3d morphable models of faces. *arXiv preprint arXiv:1903.10384*, 2019. 3

[13] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 1, 2, 3, 6, 7

[14] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020. 2, 3, 7

[15] Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. Capture, learning, and synthesis of 3d speaking styles. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10101–10111, 2019. 6, 8

[16] Matthias Eck, Tony DeRose, Tom Duchamp, Hugues Hoppe, Michael Lounsbery, and Werner Stuetzle. Multiresolution analysis of arbitrary meshes. In *Annual Conference on Computer Graphics and Interactive Techniques*, pages 173–182, 1995. 3

[17] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018. 2

[18] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics*, 39(5):1–38, 2020. 2

[19] Lijie Fan, Wenbing Huang, Chuang Gan, Junzhou Huang, and Boqing Gong. Controllable image-to-video translation: A case study on facial expression generation. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 3510–3517, 2019. 1

[20] Zhenfeng Fan, Silong Peng, and Shihong Xia. Towards fine-grained optimal 3d face dense registration: An iterative dividing and diffusing method. *International Journal of Computer Vision*, pages 1–21, June 2023. 4

[21] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2

[22] Baris Gecer, Alexandros Lattas, Stylianos Ploumpis, Jiankang Deng, Athanasios Papaioannou, Stylianos Moschoglou, and Stefanos Zafeiriou. Synthesizing coupled 3d face modalities by trunk-branch generative adversarial networks. In *European Conference on Computer Vision*, pages 415–433, 2020. 3

[23] Jiahao Geng, Tianjia Shao, Youyi Zheng, Yanlin Weng, and Kun Zhou. Warp-guided gans for single-photo facial animation. *ACM Transactions on Graphics*, 37(6):1–12, 2018. 2

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. 2, 5

[25] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 2

[26] Xianfeng Gu, Steven J Gortler, and Hugues Hoppe. Geometry images. In *Annual Conference on Computer Graphics and Interactive Techniques*, pages 355–361, 2002. 3, 4

[27] Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009. 6

[28] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[29] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2

[30] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision*, pages 172–189, 2018. 2

[31] Takashi Isobe, Xu Jia, Shuaijun Chen, Jianzhong He, Yongjie Shi, Jianzhuang Liu, Huchuan Lu, and Shengjin Wang. Multi-target domain adaptation with collaborative consistency learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8187–8196, 2021. 2

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 1, 2, 5, 6, 7

[33] Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. Large pose 3d face reconstruction from a single image via direct volumetric cnn regression. In *IEEE International Conference on Computer Vision*, pages 1031–1039, 2017. 3

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019. 1

[35] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*, 2020. 3

[36] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International Conference on Machine Learning*, pages 1857–1865, 2017. 2

[37] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *International Conference on Learning Representations*, 2015. 6

[38] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 1

[39] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716, 2016. 5, 7

[40] Hao Li, Thibaut Weise, and Mark Pauly. Example-based facial rigging. *ACM Transactions on Graphics*, 29(4):1–6, 2010. 2

[41] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2020. 3

[42] Feng Liu, Qijun Zhao, Xiaoming Liu, and Dan Zeng. Joint face alignment and 3d face reconstruction with application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):664–678, 2018. 2

[43] Feng Liu, Ronghang Zhu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Disentangling features in 3d face shapes for joint face reconstruction and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5216–5225, 2018. 2

[44] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[45] Shugao Ma, Tomas Simon, Jason Saragih, Dawei Wang, Yuecheng Li, Fernando De La Torre, and Yaser Sheikh. Pixel codec avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 64–73, 2021. 3

[46] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2

[47] Stylianos Moschoglou, Stylianos Ploumpis, Mihalis A. Nicolaou, Athanasios Papaioannou, and Stefanos Zafeiriou. 3dfacegan: Adversarial nets for 3d face representation, gen-

eration, and translation. *International Journal of Computer Vision*, 128(10):2534–2551, 2020. 3, 6, 7

[48] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *International Conference on Machine Learning*, pages 2642–2651, 2017. 2

[49] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[50] Stylianos Ploumpis, Evangelos Ververas, Eimear O'Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William AP Smith, Baris Gecer, and Stefanos Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4142–4160, 2020. 2

[51] Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *European conference on computer vision*, pages 818–833, 2018. 1

[52] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J Black. Generating 3d faces using convolutional mesh autoencoders. In *European Conference on Computer Vision*, pages 704–720, 2018. 3

[53] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021. 2

[54] Alla Sheffer, Emil Praun, Kenneth Rose, et al. Mesh parameterization methods and their applications. *Foundations and Trends in Computer Graphics and Vision*, 2(2):105–171, 2007. 4

[55] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2004–2018, 2020. 1

[56] Hao Tang, Hong Liu, Dan Xu, Philip H. S. Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*, 34(4):1972–1987, 2021. 3

[57] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3361–3371, 2021. 2

[58] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 7, 8

[59] Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 20333–20342, 2022. 6, 8

[60] Yiqiang Wu, Ruxin Wang, Mingming Gong, Jun Cheng, Zhengtao Yu, and Dapeng Tao. Adversarial uv-transformation texture estimation for 3d face aging. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 4, 6, 7

[61] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K Jain. Learning face age progression: A pyramid architecture of gans. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 31–39, 2018. 1

[62] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 601–610, 2020. 6

[63] Jiaolong Yang, Hongdong Li, Dylan Campbell, and Yunde Jia. Go-icp: A globally optimal solution to 3d icp point-set registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2241–2254, 2015. 3

[64] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiong Pan, Changgong Zhang, Feiying Ma, Xuansong Xie, and Chunyan Miao. Unbalanced feature transport for exemplar-based image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 15028–15038, 2021. 2

[65] Jian Zhao, Lin Xiong, Jianshu Li, Junliang Xing, Shuicheng Yan, and Jiashi Feng. 3d-aided dual-agent gans for unconstrained face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2380–2394, 2018. 2

[66] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. The spatially-correlative loss for various image translation tasks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16407–16417, 2021. 2

[67] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13545–13555, 2022. 2

[68] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 2, 5