Adversarial 3D Face Disentanglement of Identity and Expression

Anonymous FG2023 submission
Paper ID 96

***Abstract*— We propose a new framework to decompose 3D facial shape into identity and expression. Existing 3D face disentanglement methods assume the presence of a corresponding neutral (i.e. identity) face for each subject. Our method designs an identity discriminator to obviate this requirement. This is a binary classifier that determines if two input faces are from the same identity, and encourages the synthesised identity face to have the same identity features as the input face and to approach the 'apathy' expression. To this end, we take advantage of adversarial learning to train a PointNet-based variational auto-encoder and discriminator. Comprehensive experiments are employed on CoMA, BU3DFE, and FaceScape datasets. Results demonstrate state-of-the-art performance with the option of operating in a more versatile application setting of no known neutral ground truths. Code is available at `https://github.com/rmraaron/FaceExpDisentanglement`.**

I. INTRODUCTION

We tackle the problem of understanding a 3D facial image from the shape channel only (i.e. no color-texture) in order to obviate any ambient lighting requirements. The most immediate problem is how to disentangle 3D shape that results from a given subject identity and 3D shape that is as a result of a subject's facial expression. Such a decomposition has many applications; for example, facial identity and expression interpolations, facial expression transfer [6], [35], [36], as shown in Fig. 1, face recognition [11], [22], [23], [26], [27], [28], and facial animation [4], [7].

We aim to learn to disentangle identity and expression and to reconstruct 3D human faces, irrespective of whether neutral faces, corresponding to the identity of the expressive faces, can be accessed or not. To reach this goal, we propose an adversarial approach that combines a variational auto-encoder (VAE) [21] with an identity discriminator. For the VAE, we apply a PointNet-based [31] encoder and two decoders: the identity decoder and the expression decoder. We also employ it as our base network for the identity discriminator, which is a classifier that learns to determine whether or not a pair of 3D faces are from the same identity.

We employ the findings of Grasshof et al. [13], [14], which shows that the centre of the expression space is the point of apathy, where all face muscles are relaxed, and our identity discriminator is able to capture inherently similar features, i.e. identity features, from various expression faces. The extracted identity parts from the same person are assumed to be the apathy expression (i.e. emotionless with relaxed facial muscles). Conversely, the identity discriminator aims to make different identity representations distant from each other. The adversarial process drives our network to synthesise invariant identity faces from the same subject.

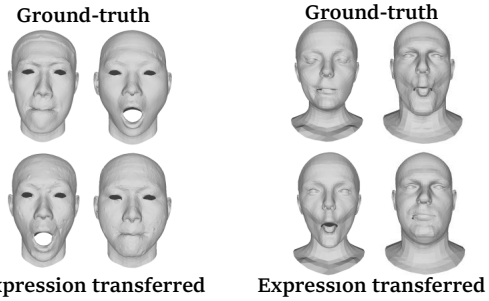


Fig. 1: Expression transfer using our disentanglement network on FaceScape data (left) and CoMA data (right).

Compared to other methods that require a corresponding neutral face for each subject, we consider the invariant, apathetic identity representations learned by the discriminator as our 'neutral', in the scenarios when we are not able to obtain ground truth neutrals. Thorough qualitative and quantitative evaluations show that our adversarial approach can disentangle identity and expression features and synthesise high quality 3D face shapes. In summary, contributions are:

- An adversarial approach to facial identity and expression disentanglement that exploits a PointNet-based VAE and discriminator.
- To the best of our knowledge, we are the first to address the scenario of unknown ground truth neutrals, leveraging the invariance of identities from same individuals and employing the apathy 'expression' as the center of expression space in order to train an end-to-end model, i.e. the identity discriminator and the VAE are trained simultaneously in an unsupervised manner.
- We compare the results of using and not using neutral ground-truths, and observe the performance of disentanglement on applications including face recognition, expression transfer and expression interpolation.
- Evaluation on publicly-available datasets demonstrates state-of-the-art results with the option of operating in a more versatile application setting of no known neutral ground truths.

II. RELATED WORK

Many recent works aim to analyse 2D and 3D images of the human face in terms of their physically-meaningful components i.e. subject identity, facial expression, surface reflectance, illumination and camera parameters. The introduction of 3D Morphable Models (3DMMs) [2] is a

notable early milestone. In subsequent years, 3D face models were developed that use more sophisticated shape morphing techniques [15], or a larger body of 3D training samples [3], or that cover the full cranium as well as the face [10] and that have articulated components [24], [30].

Several works focus on identity and expressions analysis. A statistical model [5] was employed to fit 3D faces and analysed facial identity and expression and explored their variations. Bouaziz et al. [4] combined an identity PCA model, a dynamic expression template, and a parameterized deformation model, which transformed the neutral shape to generate user-specific blendshapes.

Many nonlinear models were proposed to decouple the identity and expression features from a 3D face shape. Tewari et al. [34] presented a multi-frame video-based self-supervised training of a deep network to disentangles facial shape, appearance, expression, and illumination.

Tran and Liu [37] proposed a nonlinear 3D Morphable Model (3DMM). Liu et al. [25] also presented a framework to learn a nonlinear face model by treating 3D scans as unorganized point clouds and transforming them into shape and expression latent representations and then recovering 3D shapes. [26] utilised an encoder-decoder network to regress 3D face shapes from 2D face images and to disentangle the identity and non-identity components of 3D face shapes.

In order to construct both identity and expression 3D shape models from general 3D face datasets, we need high-performance identity-expression disentanglement, which is the aim of this work, and we now focus on prior works specifically aimed at this goal.

A. Disentangled Face Representations

Human facial expression analysis has been the focus of many studies, in which it requires an identity-agnostic expression representation. Jiang et al. [16] observe that human expressions lie in a high-dimensional manifold and that the expression manifolds of different subjects are similar. Neutral expressions, i.e. identity attributes, were set as the origin points and they proposed a nonlinear framework to decompose 3D face meshes into identity and expression attributes. Abrevaya et al. [1] introduced the use of the Generative Adversarial Network (GAN) [12] architecture for decoupling 3D facial natural factors, such as identity and expressions. Zhang et al. [41] combined a VAE with an adversarial network in order to eliminate correlations between identity and expression representations and ensure their independence. Kacem et al. [17] employed a GAN to extract expressive representations. Zhang et al. [40] modelled expressions as the deviation from the identity and extracted a deviation feature vector using a deviation learning network with a pseudo-siamese structure. Note that existing disentanglement methods take neutral expressions into consideration. We also decouple expressions from 3D faces without the requirement for corresponding neutral expression ground truths.

III. PROPOSED METHOD

In this section, we describe the details of our end-to-end method for 3D facial identity and expression disentanglement. Fig. 2 demonstrates our overall joint learning pipeline. We introduce our overall architecture in the Sec. III-A and then we explain the encoder-decoder, the identity discriminator and the loss functions.

A. Overall Architecture

We view each aligned 3D face scan $\mathbf{X}^i \in \mathbb{R}^{n\times 3}$ ($i \in [1, \cdots, m]$) as point clouds where n is the number of vertices and m is the number of input 3D face scans. (Note that we simplify $\mathbf{X}^i$ to $\mathbf{X}$ in the following.) Each instance of $\mathbf{X}$ is divided into the identity part $\mathbf{X}_{id} \in \mathbb{R}^{n\times 3}$ and the expression deformation part $\mathbf{X}_{exp} \in \mathbb{R}^{n\times 3}$. We assume identity and expressions are independent, so that the full face is the sum of the identity shape and the expression blendshape, formulated as:

$$\mathbf{X} = \mathbf{X}_{id} + \mathbf{X}_{exp}. \tag{1}$$

In our architecture, shown in Fig. 2, the whole network is designed as a GAN, where the encoder-decoder network is the 'generator' part of the network. We employ a variational encoder, based on PointNet, to learn identity and expression distributions and sample their latent vectors $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ respectively. Two decoders are used to reconstruct the identity $\bar{\mathbf{X}}_{id}$ and expression $\bar{\mathbf{X}}_{exp}$ components from corresponding latent vectors $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ and using (1), the full faces are synthesised.

Another essential part of the GAN framework is the discriminator and we propose an identity discriminator. The input of this discriminator is a face shape *pair* containing a 3D face $\mathbf{X}$ ($\mathbf{X}^i$) and its predicted identity shape $\bar{\mathbf{X}}_{id}$ ($\bar{\mathbf{X}}^i_{id}$) from the identity decoder or another 3D face shape $\mathbf{X}^j$ with the same identity. This discriminator is trained to distinguish a 'real' face shape pair (i.e. same identity) from a 'fake' pair (i.e. different identity). When jointly training the end-to-end model, the original face shape pairs with same identities $(\mathbf{X}^i, \mathbf{X}^j)$ are considered as real samples, and those pairs that include predicted identity shapes from the identity decoder $(\mathbf{X}^i, \bar{\mathbf{X}}_{id})$ are considered as fake. Thus, the generator is encouraged to learn an intrinsic identity latent distribution in the process of adversarial learning.

B. Variational Encoder-Decoder Network

Although we aim to disentangle 3D face identity shapes and expressions, 3D face reconstruction is also considered. We employ a VAE network, in which the encoder is used to predict distributions of latent representations from input point clouds and the decoders are used to recover these 3D face shapes. To better decouple identity and expressions, the encoder outputs distributions for identities and expressions separately, and two decoder branches, i.e. ID decoder and EXP decoder, receive their corresponding sampled representations and reconstruct 3D identity face shapes and expression blendshapes individually.

The VAE models the probability $P(\mathbf{X})$ of the input 3D face shapes and we assume that 3D face shapes are determined by latent features $\mathbf{z}_{id}$ and $\mathbf{z}_{exp}$ representing identity and expression respectively. This generative model estimates

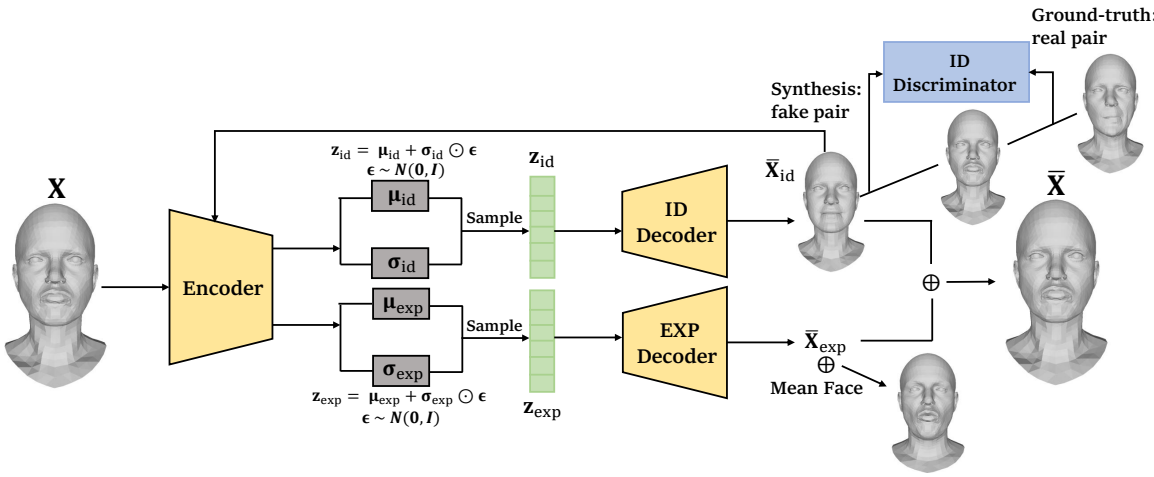


Fig. 2: A framework of 3D face identity and expression disentanglement. This joint learning network includes a variational encoder-decoder part for 3D face reconstruction and a discriminator to enforce the same identity in an adversarial manner

parameters that maximize the likelihood of 3D face identities and expressions, as follows:

$$p_\theta(\mathbf{X}_{id}) = \int p_\theta(\mathbf{z}_{id})\, p_\theta(\mathbf{X}_{id}|\mathbf{z}_{id})\, d\mathbf{z}, \quad (2)$$

$$p_{\theta'}(\mathbf{X}_{exp}) = \int p_{\theta'}(\mathbf{z}_{exp})\, p_{\theta'}(\mathbf{X}_{exp}|\mathbf{z}_{exp})\, d\mathbf{z}, \quad (3)$$

where $p_\theta(\mathbf{X}_{id}|\mathbf{z}_{id})$ and $p_{\theta'}(\mathbf{X}_{exp}|\mathbf{z}_{exp})$ represent the identity decoder and expression decoder respectively. We assume a unit Gaussian distribution for the prior distributions $p_\theta(\mathbf{z}_{id})$ and $p_{\theta'}(\mathbf{z}_{exp})$.

Due to the intractable posterior $p_\theta(\mathbf{z}_{id}|\mathbf{X}id)$, the distribution $q_\phi(\mathbf{z}_{id}|\mathbf{X}_{id})$ is defined in the identity encoder to approximate $p_\theta(\mathbf{z}_{id}|\mathbf{X}_{id})$. We use the Kullback-Leibler (KL) divergence term $D_{KL}(q_\phi(\mathbf{z}_{id}|\mathbf{X}_{id})||p_\theta(\mathbf{z}_{id}|\mathbf{X}_{id}))$ to minimize their difference. Similarly for $p_{\theta'}(\mathbf{z}_{exp}|\mathbf{X}_{exp})$.

The VAE aims to maximize the log-likelihood of 3D facial identities and expressions, taking the identity branch as an example:

$$\log p_\theta(\mathbf{X}_{id}) = \mathbf{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}_{id}|\mathbf{X}_{id})}[\log p_\theta(\mathbf{X}_{id})] \geq \text{ELBO}, \quad (4)$$

where ELBO is defined as the following expectation:

$$\mathbf{E}_{\mathbf{z}}[\log p_\theta(\mathbf{X}_{id}|\mathbf{z}_{id})] - D_{KL}(q_\phi(\mathbf{z}_{id}|\mathbf{X}_{id})||p_\theta(\mathbf{z}_{id})). \quad (5)$$

Thus, the VAE is assumed to estimate parameters θ and ϕ to maximize the ELBO (Evidence Lower Bound) in (4) and (5). In other words, negative ELBO is considered as one of the loss function terms in our network.

C. Adversarial Training

Face identity shapes and expression blendshapes are sampled from the $p(\mathbf{X}|\mathbf{z})$ distributions and to promote better decoupling of identity shapes from expressions, an adversarial training process is employed.

Our proposed identity discriminator $\mathbf{D}_{ID}$ is trained to distinguish between real and fake samples. Additionally, we input a 3D face shape pair into the identity discriminator that decides whether the input pair has the same identity shape.

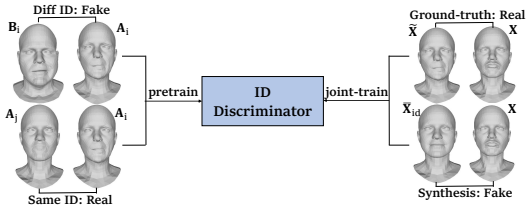
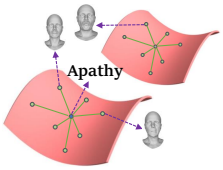
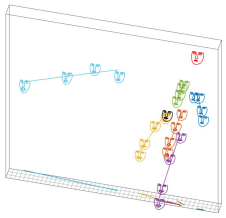


Fig. 3: The pre-train input pairs and the joint end-to-end train input pairs of the identity (ID) discriminator



(a) 3D facial shape space



(b) Expression affine subspace [13] (© 2021 IEEE)

Fig. 4: Illustrations of 3D facial shape space and expression affine subspace. In the second sub-figure, there is a special point indicated by the top-right red face: the point of apathy

For instance, if we input a face scan pair $\mathbf{A}_i$ and $\mathbf{A}_j$ with the same identity $\mathbf{A}_{id}$, this identity discriminator is expected to classify this pair into the real class (note that the subscripts (i, j) index expressions). Otherwise, a pair $\mathbf{A}_i$ and $\mathbf{B}_i$ should be classified into the fake class due to one with the identity $\mathbf{A}_{id}$ and another with the identity $\mathbf{B}_{id}$, as illustrated in Fig. 3.

In facial expression analysis [8], [9], [33], the latent variables that represent identity and facial expression, lie on a manifold in high dimensional space, as illustrated in Fig. 4a. Stella et al. proposed that the point of apathy is the centre of expressions [13], [14], and expressions trajectories obtained by varying the strength of human emotion originate from this point, as shown in Fig. 4b. Based on these observations, the implicit connection of faces with the same identity, notwith-

standing various expressions, is the apathetic expression. Our adversarial process encourages common information from a face shape pair with the same identity to be retained. If we only compare the smile expression with the surprised expression from the same individual, not only their same identity but their similar expression deformations, e.g. open mouth, will be recorded by the discriminator. However, there are several shape pairs from the same identity and their expressions are distributed in divergent directions that intersect at the point of apathy - so the discriminator will ultimately retain all pairs' common information - *apathy*. Thus, the identity discriminator has the ability to capture similar latent features, i.e. identity features, between pairs belonging to same subject, and to enforce these features to be close to the their *apathy* faces.

In a GAN framework, a generator and discriminators are trained adversarially. In our network, the ideal situation is that synthesised face identity shapes from the VAE can fool the discriminator. In other words, the predicted neutral face shapes are close enough to the corresponding facial identities to make the discriminator believe that they belong to the 'real' class. During the adversarial process, the discriminator takes advantage of a loss function that enables the identity decoder distribution $p\left(\mathbf{X}_{id}|\mathbf{z}_{id}\right)$ to approximate to the face identity distribution $p\left(\mathbf{X}_{id}\right)$. The loss used to jointly train the generator and discriminator is:

$$J_{adv} = \max_{\theta_d}\Big[\mathbf{E}_{(\widetilde{\mathbf{X}},\mathbf{X})\sim p_{same}}\left[\log\left(\mathbf{D}_{ID}\left(\widetilde{\mathbf{X}},\mathbf{X};\theta_d\right)\right)\right] + \mathbf{E}_{\mathbf{X}\sim p_{data}}\left[\log\left(1-\mathbf{D}_{ID}\left(\mathbf{G}\left(\mathbf{X}\right),\mathbf{X};\theta_d\right)\right)\right]\Big], \quad (6)$$

where p_{same} is the distribution of all same-identity 3D face pairs. The sampled pair from p_{same}, i.e. $\left(\widetilde{\mathbf{X}},\mathbf{X}\right)$ in (6), is equivalent to sampling from the original dataset twice, such that the two sampled faces have the same identity. $\mathbf{G}\left(\mathbf{X}\right)$ is the combination of the encoder and the decoder, and its output is the synthesised identity shape $\bar{\mathbf{X}}_{id}$.

D. End-to-end Loss Function Terms

We define five components in our loss function that is required to train our end-to-end network for 3D face reconstruction and identity-expression disentanglement. The overall loss function is:

$$L_{total} = \lambda_1 L_{recon} + \lambda_2 L_{KL} + \lambda_3 L_{\mathbf{D}_{ID}} + \lambda_4 L_{neu} + \lambda_5 L_{lap}, \quad (7)$$

where λ_{1-5} are hyperparameters to balance these five losses. The L_{recon} is the Mean Squared Error (MSE) for 3D face reconstruction; the L_{KL} loss is the ELBO term from (5). We use two KL losses, one for the identity part and another for the expression part, to constrain the posterior distribution close to the unit Gaussian distribution $\mathcal{N}(0,I)$. The $L_{\mathbf{D}_{ID}}$ in (8) is simplified from the (6), by using a cross entropy.

$$L_{\mathbf{D}_{ID}} = -\Big[y\log\left(\mathbf{D}_{ID}\left(\widetilde{\mathbf{X}},\mathbf{X}\right)\right) + (1-y)\log\left(\mathbf{D}_{ID}\left(\bar{\mathbf{X}}_{id},\mathbf{X}\right)\right)\Big], \quad (8)$$

where y is the label (1 for the 3D face pair $\left(\widetilde{\mathbf{X}},\mathbf{X}\right)$ sampled from ground truth data and 0 for the 3D face pair including the predicted identity shape $\bar{\mathbf{X}}_{id}$). After facial identity shapes $\bar{\mathbf{X}}_{id}$ are predicted by the identity decoder, $\bar{\mathbf{X}}_{id}$ is fed into the encoder again, and the associated identity latent vector $\bar{\mathbf{z}}_{id}$ is sampled. L_{neu} is the L1 loss for $\mathbf{z}_{id}$ and $\bar{\mathbf{z}}_{id}$. To minimise the mean curvature and make 3D faces smooth, we employ a Laplacian regularization loss L_{lap} [18] that is written as $L_{lap} = \|L\mathbf{X}\|_2$, where L is the discrete Laplace-Beltrami operator.

IV. EVALUATION

In this section, we present an experimental evaluation of our proposed reconstruction and disentanglement method. Firstly, datasets, implementation details, and evaluation metrics are introduced. We compare our methods (both with and without neutral ground-truths) against baselines on three public datasets. We also perform ablation studies to analyse the benefits of the components in our architecture design. To demonstrate the utility and effectiveness of our approach, we show the results of several applications, including: expression transfer, expression interpolation and face recognition (latter in Supplementary Material).

A. Datasets Employed

The three datasets used are given below. In BU3DFE and CoMA the ratio of training set size to test set size is 9:1, and in FaceScape the ratio is 7:3.

CoMA dataset [32] contains motion sequences of 20,466 meshes from 12 different individuals. Each subject performs 12 extreme, asymmetric facial expressions. We follow [32] and divide these meshes into a training and test set, so that the sequences are fixed in alphabetical order and we take 10 frames from every 100 frames as test samples. There are 18,422 and 2040 meshes in the training set and test set respectively.

BU3DFE [39] includes 100 subjects with 2500 3D facial scans, and each subject is asked to perform seven expressions. With the exception of the neutral expression, each of the six other expressions includes four levels of intensity. We follow [41] and select the first 10 subjects as the test set and the rest are used for training. There are 2247 meshes in the training set and 250 meshes in the test set. The test identities are unseen in training.

FaceScape [38] contains 847 subjects and each subject performs 20 expressions. We randomly select 30% of the subjects as the test set and the rest are used for training. There are 11,812 and 5055 meshes in the training and test set respectively. The test identities are unseen in training.

B. Implementation Details

In the FaceScape dataset, there are 26,317 vertices and 52,261 faces per subject. This is prohibitive in terms of GPU memory and time when training and so we simplify meshes using a quadric-based edge collapse strategy after which each mesh includes 4547 vertices and 8999 faces.

We pretrain a PointNet-based network as the identity discriminator. As depicted in Fig. 3, we sample a 3D face scan each time from the training dataset, and the specific

input and the sampled face shape constitute a pair. If the subject identity of a pair is the same, the label is set as "True", otherwise the label is "False".

We adopt the pretrained identity discriminator as an initialisation in the joint end-to-end training, and explore alternative pairs to train the encoder-decoder and the identity discriminator together. The new "True" pair represents the input ground truth 3D face shape and a sampled one with the same identity, whereas the predicted identity shape from identity decoder and the original one are regarded as a "False" pair.

For a fair comparison, the PointNet-based encoder takes four identity latent dimensions and four expression latent dimensions, which is the same as the compared methods, in the CoMA dataset. For BU3DFE we use 40 dimensions for each latent vector and for FaceScape we use 64.

We implement the network with PyTorch [29] and run it on an NVIDIA A40 system. We pretrain the identity discriminator using a batch size of 32 and 50, 100, 100 epochs on CoMA, BU3DFE, and FaceScape respectively. End-to-end networks are trained using the Adam optimiser [20] with the learning rate at 1×10^{-4}, and a learning rate decay is set as 0.7 every 50 epochs.

Different hyperparameters for different datasets are explored to balance each loss, including λ_1 being set as 250 on BU3DFE and 5000 on CoMA and FaceScape, and λ_3 as 5×10^{-4} on CoMA as well as FaceScape and 1×10^{-3} on BU3DFE. In FaceScape and CoMA, we only use the Laplacian loss to make predicted identity faces smooth when neutral ground-truths are not available. We train the joint end-to-end network for 280 epochs and a batch size of 8 on BU3DFE, 280 epochs and a batch size of 32 on FaceScape, and 300 epochs and a batch size of 32 on CoMA. We conduct every leave-one-out experiment three times and report their average results.

C. Evaluation Metrics

We adopt the evaluation metrics used in [16], [17], [41], i.e. both reconstruction and disentanglement metrics. Our system is based on point clouds, so the average vertex distance between synthesised 3D face shapes $\bar{\mathbf{X}}$ and original 3D face shapes $\mathbf{X}$ is considered as the reconstruction error E_{rec} in (9):

$$E_{rec}(\mathbf{X}, \bar{\mathbf{X}}) = \frac{1}{n} \sum_{i=1}^{n} \left\| \mathbf{X}_i - \bar{\mathbf{X}}_i \right\|_2, \tag{9}$$

where n is the number of vertices of $\mathbf{X}$. We report the mean and the median of the average vertex distance.

The standard deviation of reconstructed identity shapes from 3D faces with the same identity is assumed as the disentanglement error E_{dis}. This is designed to evaluate the disentanglement. Given a test set containing raw 3D faces with various identity shapes that are represented as $\mathbf{A}^{id}$, $\mathbf{B}^{id}$, $\cdots$, $\mathbf{N}^{id}$, the raw 3D faces owning the same identity $\mathbf{A}^{id}$ and different expressions are denoted as $\mathbf{A}_i$. The disentanglement error E_{dis} is computed as:

$$E_{dis} = \sigma \left(\left\| \mathbf{A}_i^{id} - \mathbf{A}_{mean}^{id} \right\|_2 \right), \tag{10}$$

where $\mathbf{A}_{mean}^{id}$ is the mean of predicted identity shapes $\mathbf{A}_i^{id}$ from $\mathbf{A}_i$ ($i \in [1, \cdots, k]$), and k is the number of expression types. We do not quantitatively evaluate the predicted expression shape because we consider that different people express the same expressions in slightly different ways, which is affected by their anatomy.

The average vertex distance AVD_{neu} of identity shapes is used to evaluate the reconstruction and disentanglement process as well.

D. Comparison with Recent Literature

We compare our work with five state-of-the-art 3D face disentanglement methods, FLAME [24], Jiang et al. [16], and DI-MeshEncoder [41] on CoMA and BU3DFE, and Kacem et al. [17] and Convolutional Mesh Autoencoder (Conv-MeshAE) [32] on FaceScape. We carefully report results from [41] since we have the same training and test set. The FLAME model is factored in the sense that it separates the representation of identity, pose, and facial expression. It includes a learned shape space of identity variations and expression blendshapes to capture non-rigid deformations of faces. Jiang et al. and DI-MeshEncoder adopt a Graph Convolutional Network (GCN) based autoencoder to reconstruct 3D face shapes and decouple identity and expression attributes. Kacem et al. uses a GCN network and a discriminator for expression neutralisation and face recognition. Conv-MeshAE proposes a GCN architecture to represent 3D face shapes into non-linear latent space. Although this is a 3D reconstruction method instead of disentangling identity and expression shapes, we follow [17] to consider pairs of expressive and neutral faces as its input and ground-truths, respectively. We use a widely-adopted autoencoder structure based on PointNet that is unlike these four methods and, furthermore, neutral ground truths are not required in our method. Our discriminator is employed on raw data. This is different from [41] who use a discriminator to enforce independence of two distributions, which is based on Kim and Mnih's PMLR 2018 work [19]. Our work is also different from [17], who use a discriminator in the latent space to learn a valid translation from expressive to neutral representations.

E. Results and Discussions

TABLE I: Disentanglement (E_{dis}) and reconstruction results (E_{rec}) on CoMA. Compared with FLAME, Jiang et al., DI-MeshEnc and Conv-MeshAE. All errors are in millimeters

Methods	E_{dis}		E_{rec}	
	mean	med	mean $\pm$ std	med
FLAME [24]	0.599	0.591	1.451 $\pm$ 1.649	0.871
Jiang et al. [16]	0.064	0.062	1.413 $\pm$ 1.639	1.017
DI-MeshEncoder [41]	0.019	0.020	0.665 $\pm$ 0.748	**0.434**
Conv-MeshAE [32]	0.313	0.317	—	—
Ours	0.176	0.180	0.783 $\pm$ 0.225	0.772
Ours+ne-gt	**0.014**	**0.013**	**0.651** $\pm$ **0.208**	0.625

The quantitative results on CoMA and BU3DFE that are compared with FLAME, Jiang et al., DI-MeshEncoder and

TABLE II: Disentanglement and reconstruction results on BU3DFE. Compared with FLAME, Jiang et al., DI-MeshEnc and Conv-MeshAE. All errors are in millimeters

Methods	E_{dis}		E_{rec}	
	mean	med	mean ± std	med
FLAME [24]	0.600	0.632	2.596 ± 2.055	2.055
Jiang et al. [16]	0.611	0.590	2.054 ± 1.199	1.814
DI-MeshEnc [41]	0.361	**0.327**	1.551 ± 0.924	1.375
Conv-MeshAE [32]	0.361	0.377	—	—
Ours	0.443	0.439	**1.421 ± 0.412**	**1.306**
Ours+ne-gt	**0.348**	0.339	1.500 ± 0.423	1.467

TABLE III: Average vertex distance of identity shapes AVD_{neu} and disentanglement results E_{dis} on FaceScape. Compared with [17] and [32]. All errors are in millimeters

Methods	AVD_{neu}		E_{dis}	
	mean ± std	median	mean	med
Kacem et al. [17]	2.02 ± —	—	—	—
Conv-MeshAE [32]	2.00 ± 0.52	1.90	0.64	0.62
Ours	3.11 ± 0.92	2.96	0.77	0.76
Ours+ne-gt	**1.93** ± 0.61	**1.82**	**0.57**	**0.55**

Conv-MeshAE are given in Tab. I and Tab. II respectively. E_{rec} of Conv-MeshAE is missing because its generated shapes are identities instead of original input faces. We also report $\textbf{AVD}_{neu}$ that is compared with Conv-MeshAE on BU3DFE and CoMA in the Supplementary Material. We only compare with Conv-MeshAE and Kacem et al. on FaceScape with the results listed in Tab. III, since the FaceScape dataset was published recently and there are very few disentanglement experiments on it. Kacem et al. adopts different training and test set splits and predicts neutral shapes of unseen identities on the CoMA dataset. We also conduct experiments with this split scheme and report results in the Supplementary Material.

The "Ours" in these tables means that our method *does not access the neutral ground-truths in end-to-end training*, which fits to some real-world scenarios where corresponding identity shapes are not available. The "Ours+ne-gt" denotes that we use neutral faces as ground-truths, as is the case with all the methods that we compare with.

From Tab. I and Tab. II, we observe that we achieve improvements on E_{dis} in CoMA and E_{rec} in BU3DFE. From Tab. II, we see one of our methods with the best E_{dis} and the other with the best E_{rec}. The reason is that we employ a GAN network and there is a trade-off between reconstruction and disentanglement performance. Unsurprisingly, disentanglement performance (represented by E_{dis}) drops when neutral ground-truths are not accessed in our method, especially in the CoMA dataset. This lower E_{dis} is a result of the small number of identities. On average, more than 1500 meshes have the same identity, so there are many mesh-pairs of the same identity. The variance of retained common information is larger than those with strong supervision or those with fewer pairings in the discriminator training. Tab. III shows that our method also has strong performance on the FaceScape dataset.

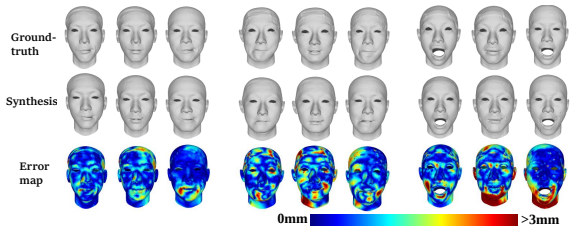


Fig. 5: Results of unseen 3D face disentanglement using neutrals on FaceScape - from left to right group: the best, the average, and the worst. Each group has three faces: full face, neutral and expression

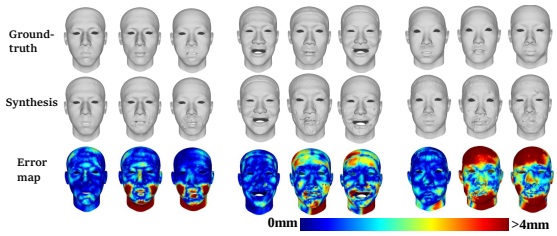


Fig. 6: Results of unseen 3D face disentanglement without neutrals on FaceScape. The same arrangement as Fig. 5

In Fig. 5, we select some representatively unseen identity results of 3D face reconstructions and disentanglement using neutral ground-truths on FaceScape. They are divided into three groups to show our best, average and worst performance (based on quantitative metrics). In each group, the first row consists of the ground truth full face, neutral, and expression. The second row consists of corresponding prediction results. We show the error heat maps in the third row. Fig. 6 reports results of 3D face disentanglement and reconstruction without ground truth neutrals on FaceScape. As shown in Fig. 7 and Fig. 8, the predicted neutral faces have extremely low error when using neutral ground-truths on CoMA, and expression predictions perform slightly worse than identity parts. The unseen identity results of BU3DFE are illustrated in Fig. 9 and Fig. 10. In the worst case without neutral ground-truths of BU3DFE and FaceScape, the predicted identity shapes have slight expressions (open / left mouth), especially when expression shapes are exaggerated.

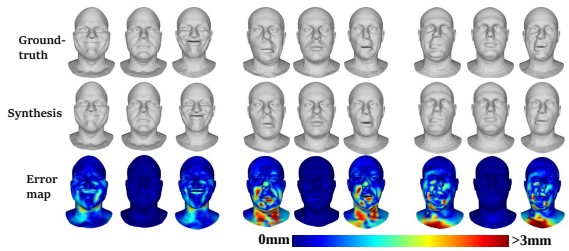


Fig. 7: Results of 3D face disentanglement using neutrals on CoMA. The same arrangement as Fig. 5

TABLE IV: Comparison results of E_{rec}, E_{dis} and AVD$_{neu}$ on CoMA, BU3DFE and FaceScape. The '−ne-gt' means our methods without neutral ground-truths and '−id-dis' means our method without the identity discriminator. Conversely, the '+ne-gt' and '+id-dis' mean our method with neutrals and the identity discriminator respectively. All errors are in millimeters

Methods		Dataset	AVD$_{neu}$		E_{dis}		E_{rec}	
			mean ± std	median	mean	median	mean ± std	median
−ne-gt	−id-dis	CoMA	2.775 ± 0.948	2.509	1.439	1.443	**0.686 ± 0.190**	0.674
	+id-dis		**1.528 ± 0.675** ↓	**1.268** ↓	**0.176** ↓	**0.180** ↓	0.783 ± 0.225	**0.651**
	−id-dis	BU3DFE	4.108 ± 1.246	3.958	1.211	1.144	1.469 ± 0.405	1.359
	+id-dis		**2.429 ± 0.667** ↓	**2.283** ↓	**0.443** ↓	**0.439** ↓	**1.421 ± 0.412** ↓	**1.306** ↓
	−id-dis	FaceScape	12.020 ± 0.514	11.876	1.791	1.795	1.187 ± 0.300	1.138
	+id-dis		**3.112 ± 0.916** ↓	**2.957** ↓	**0.765** ↓	**0.758** ↓	**1.157 ± 0.286** ↓	**1.109** ↓
+ne-gt	−id-dis	CoMA	0.071 ± 0.012	0.070	0.016	0.015	0.669 ± 0.213	0.647
	+id-dis		**0.065 ± 0.012** ↓	**0.063** ↓	**0.014** ↓	**0.013** ↓	**0.651** ↓ **± 0.208** ↓	**0.625** ↓
	−id-dis	BU3DFE	**1.885 ± 0.459**	**1.733**	**0.345**	**0.337**	1.509 ± 0.427	**1.382**
	+id-dis		1.894 ± 0.430	1.764	0.348	0.339	**1.500 ± 0.423** ↓	1.404
	−id-dis	FaceScape	1.927 ± 0.617	1.821	0.582	0.563	1.393 ± 0.379	1.330
	+id-dis		**1.927 ± 0.610** ↓	**1.815** ↓	**0.569** ↓	**0.551** ↓	**1.370 ± 0.369** ↓	**1.307** ↓

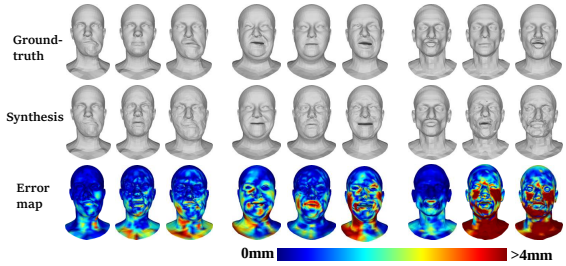


Fig. 8: Results of 3D face disentanglement without neutrals on CoMA. The same arrangement as Fig. 5

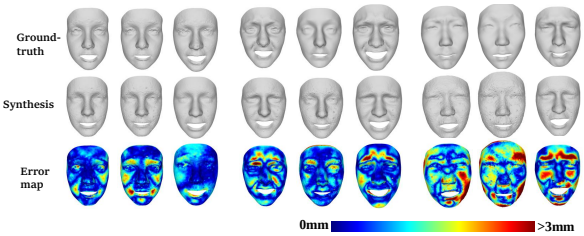


Fig. 9: Results of unseen 3D face disentanglement using neutrals on BU3DFE. The same arrangement as Fig. 5

F. Ablation Study

We now study the effectiveness of our discriminator. From Tab. IV, we can observe that our discriminator greatly improves disentanglement performance when we cannot access neutral ground-truths. For example, our '+id-dis' outperforms the '−id-dis' on FaceScape without ground-truth neutrals by around 75% of AVD$_{neu}$ and 57% of E_{dis} (decreasing from 12.020 to 3.112 and from 1.791 to 0.765 respectively). The same effectiveness is qualitatively depicted in Fig. 11. The improvements of disentanglement with neutral ground-truths are not as significant as the case with unavailable ground-truths. Note that using ground-truth neutrals is a strong supervised training process whereas, in contrast, the VAE and discriminator learn identity representations adversarially in an weakly-supervised process. Thus, when ground-truth neutrals (strong supervised process) works, the effectiveness of weakly-supervised process is not obvious.

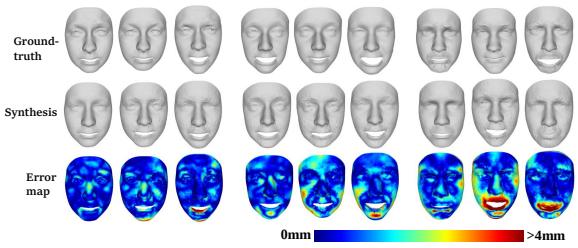


Fig. 10: Results of unseen 3D face disentanglement without neutrals on BU3DFE. The same arrangement as Fig. 5

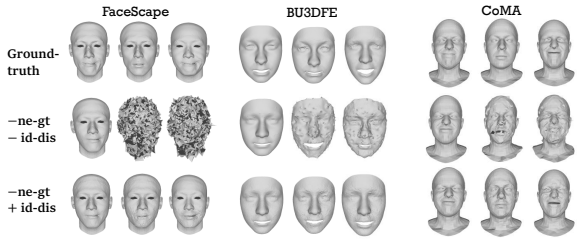


Fig. 11: Comparisons on using/not-using the identity discriminator when neutral ground truths are unknown on three datasets: FaceScape, BU3DFE and CoMA. Each datase has three faces: full face, neutral and expression

In addition, some reconstruction results are compromised a small degree because of the adversarial learning.

G. Applications

We apply our network in expression transfer, identity and expression interpolation, and face recognition. Taking CoMA and FaceScape, we randomly select two subjects with different expressions in the test set and transfer their expression latent representations, as shown in Fig. 1. We display disentangled identity and expression interpolations in the Supplementary Material. We implement face recognition on FaceScape and BU3DFE, since there are only 12 individuals on CoMA, and we compare it with Kacem et al. and Conv-MeshAE. The results are also published in the Supplementary Material and a very similar to each other.

V. Conclusions

We proposed a method employing a VAE and a discriminator for disentangling 3D face identities and expressions. To learn identity representations, we use pairs of 3D faces to train an identity discriminator, which is forced to capture identity features of the same subjects only. This particularly improves the performance in the situations where neutral expressions are not available. Additionally, the joint end-to-end learning of the encoder-decoder network and the identity discriminator helps reconstruct 3D faces. We perform evaluations on CoMA, FaceScape and BU3DFE, showing the high effectiveness of our network for 3D face reconstruction and identity/expression disentanglement.

References

[1] V. F. Abrevaya, A. Boukhayma, S. Wuhrer, and E. Boyer. A generative 3d facial model by adversarial training. 2019.

[2] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999.

[3] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou. Large scale 3d morphable models. *International Journal of Computer Vision*, 126:233–254, 2018.

[4] S. Bouaziz, Y. Wang, and M. Pauly. Online modeling for realtime facial animation. *ACM Transactions on Graphics (ToG)*, 32(4):1–10, 2013.

[5] A. Brunton, T. Bolkart, and S. Wuhrer. Multilinear wavelets: A statistical shape space for human faces. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.

[6] C. Cao, Q. Hou, and K. Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on graphics (TOG)*, 33(4):1–10, 2014.

[7] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):1–10, 2013.

[8] Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006.

[9] Y. Chang, C. Hu, and M. A. Turk. Manifold of facial expression. In *AMFG*, pages 28–35, 2003.

[10] H. Dai, N. Pears, W. Smith, and C. Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128(2):547–571, 2019.

[11] H. Drira, B. B. Amor, A. Srivastava, M. Daoudi, and R. Slama. 3d face recognition under expressions, occlusions, and pose variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2270–2283, 2013.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[13] S. Grasshof, H. Ackermann, S. S. Brandt, and J. Ostermann. Multilinear modelling of faces and expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3540–3554, 2021.

[14] S. Graßhof, H. Ackermann, S. S. Brandt, and J. Ostermann. Apathy is the root of all expressions. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, pages 658–665, 2017.

[15] IEEE. *A 3D Face Model for Pose and Illumination Invariant Face Recognition*, Genova, Italy, 2009.

[16] Z.-H. Jiang, Q. Wu, K. Chen, and J. Zhang. Disentangled representation learning for 3d face shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11957–11966, 2019.

[17] A. Kacem, K. Cherenkova, and D. Aouada. Disentangled face identity representations for joint 3d face recognition and expression neutralisation. *arXiv preprint arXiv:2104.10273*, 2021.

[18] A. Kanazawa, S. Tulsiani, A. A. Efros, and J. Malik. Learning category-specific mesh reconstruction from image collections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–386, 2018.

[19] H. Kim and A. Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[21] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[22] Y. Lei, M. Bennamoun, M. Hayat, and Y. Guo. An efficient 3d face recognition approach using local geometrical signatures. *Pattern Recognition*, 47(2):509–524, 2014.

[23] H. Li, D. Huang, J.-M. Morvan, Y. Wang, and L. Chen. Towards 3d face recognition in the real: a registration-free approach using fine-grained matching of 3d keypoint descriptors. *International Journal of Computer Vision*, 113(2):128–142, 2015.

[24] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

[25] F. Liu, L. Tran, and X. Liu. 3d face modeling from diverse raw scan data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9408–9418, 2019.

[26] F. Liu, R. Zhu, D. Zeng, Q. Zhao, and X. Liu. Disentangling features in 3d face shapes for joint face reconstruction and recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5216–5225, 2018.

[27] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 212–220, 2017.

[28] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[30] S. Ploumpis, E. Ververas, E. O'Sullivan, S. Moschoglou, H. Wang, N. Pears, W. Smith, B. Gecer, and S. P. Zafeiriou. Towards a complete 3D morphable model of the human head. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[31] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[32] A. Ranjan, T. Bolkart, S. Sanyal, and M. J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 725–741, 2018.

[33] C. Shan, S. Gong, and P. W. McOwan. Appearance manifold of facial expression. In *International Workshop on Human-Computer Interaction*, pages 221–230. Springer, 2005.

[34] A. Tewari, F. Bernard, P. Garrido, G. Bharaj, M. Elgharib, H.-P. Seidel, P. Pérez, M. Zollhofer, and C. Theobalt. Fml: Face model learning from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10812–10822, 2019.

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

[36] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[37] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018.

[38] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[39] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*, pages 211–216. IEEE, 2006.

[40] W. Zhang, X. Ji, K. Chen, Y. Ding, and C. Fan. Learning a facial expression embedding disentangled from identity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6759–6768, 2021.

[41] Z. Zhang, C. Yu, H. Li, J. Sun, and F. Liu. Learning distribution independent latent representation for 3d face disentanglement. In *2020 International Conference on 3D Vision (3DV)*, pages 848–857. IEEE, 2020.