Disentangled Representation Transformer Network for 3D Face Reconstruction and Robust Dense Alignment

Authors: Xiangzheng Li, Xiaofei Li, Jia Deng, Li Xiangzheng

Date: 2023-03-02T00:00:00+00:00

Abstract

Unlike existing methods, we propose a disentangled representation Transformer network (DRTN) for 3D dense face alignment and reconstruction. Unlike traditional 3DMM-based methods, in which the target parameters (i.e., shape, expression, and pose parameters) are individually estimated without considering their direct influences on each other, although they are jointly optimized, our DRTN aims by learning···

Full Text

Disentangled Representation Transformer Network for 3D Face Reconstruction and Robust Dense Alignment

Xiangzheng Li, Xiaofei Li, Jia Deng, Xiangzheng Li*
School of Mathematics and Computer Science, Ningxia Normal University, Guyuan, 756000, China
*Corresponding author
Email: 82021011@nxnu.edu.cn

Abstract—In this paper, we propose a Disentangled Representation Transformer Network (DRTN) for 3D dense face alignment and reconstruction. Unlike traditional 3DMM-based approaches that individually estimate target parameters (shape, expression, and pose) without considering their direct mutual influences—although they may be jointly optimized—our DRTN aims to enhance the semantic representation of facial attributes by learning the correlations among different 3D facial attribute parameters. To achieve this, we present a novel strategy for designing disentangled 3D face attribute representation, which decomposes given facial attributes into identity, expression, and pose components. Specifically, the estimation of 3D face parameters in the regression network depends on the correlation with other face attribute parameters rather than

being independent. The identity branch reinforces the learning of expression and pose attributes by preserving the overall facial geometry structure while keeping identity intact. Accordingly, the expression and pose branches maintain the consistency of their respective attributes, helping refine reconstruction and alignment of facial details in large poses primarily through coupling with other facial attribute parameters. Extensive qualitative and quantitative experimental results on widely-evaluated benchmarking datasets demonstrate that our approach achieves competitive performance compared to state-of-the-art methods.

Index Terms—Face alignment and reconstruction, convolutional neural networks, disentangled representation learning, feature interaction, biometrics.

I. INTRODUCTION

3D face reconstruction is an essential topic in computer vision and graphics. The human face is one of the most discriminative parts of the human body, with facial features and contours containing abundant attributes and semantic information. Single-view 3D face reconstruction aims to recover a complete facial geometry shape from a given single-view image, which plays a significant role in many visual analysis applications such as face recognition [**?**], face verification [**?**], facial expression analysis [**?**], and facial animation [**?**].

However, many face images are affected by shooting angles and surrounding environments. Problems such as partial facial occlusion and blurring occur, resulting in distortion of reconstructed 3D faces. Therefore, how to perform accurate 3D face reconstruction and facial landmark alignment under large face poses, partial occlusion, and insufficient illumination remains an unsolved problem.

Conventional methods are mainly based on optimization algorithms, e.g., Iterative Closest Point [**?**], Shape from Shading [**?**, **?**], and Photometric Stereo [**?**]. Nevertheless, these face reconstruction techniques suffer from problems of local optima, short model initialization, and high optimization complexity, making the 3D face reconstruction process complex and costly.

With the rise of deep learning in recent years, CNN-based regression methods [**?**, **?**] have emerged and achieved remarkable success in 3D face reconstruction and dense alignment. It is very challenging to reconstruct 3D face shapes from 2D images without prior knowledge, primarily because 2D data does not convey clear depth information. A common approach to solving single-view 3D face reconstruction involves using heat maps, which are regressed with input RGB images by learning 3D face depth values through a residual network. A video-based 3D cascade regression method has been developed by Jeni et al. [**?**]. This method generates a dense 3D shape in real-time from an input 2D face image by estimating the position of a dense set of markers and their visibil-

ity, then achieves dense 3D face alignment by fitting a partial 3D model. Xin Ning et al. [**?**] proposed a real-time 3D face alignment method that uses an encoder-decoder network with efficient convolutional layers to enhance information transfer between different resolutions in the encoding and decoding stages, achieving advanced performance.

Currently, several works perform face alignment by fitting 3D Morphable Models (3DMM) to 2D facial images. [**?**] fits a 3DMM using a single CNN in an iterative manner, where the CNN augments the input channel with represented shape features in each iteration. [**?**] uses multi-constraint training CNNs to estimate 3DMM parameters and provides very dense 3D alignment. Lei Jiang et al. [**?**] proposed a dual-attention mechanism and a practical end-to-end 3D face alignment framework, constructing a stable network model through deep separable convolution, densely connected convolution, and a lightweight channel attention mechanism. This model-based 3D reconstruction method can easily accomplish the task of 3D face alignment. However, these methods only use deep networks to directly learn the relationship between 2D input images and 3D models without considering the integrity of correspondence between face attributes. To address this, our model decomposes face attributes through a unified deep learning architecture and automatically extracts useful feature information directly from pixels using complementary information between face attributes. This approach infers face contours in invisible regions and improves face alignment in large poses.

B. 3D Face Reconstruction

Initially, 3D face reconstruction was mainly used in medical applications for human head diagnosis, typically done by scanning faces with a 3D scanner [**?**, **?**] to obtain shape, structure, and texture information for reconstructing a complete 3D face. Although reconstructing 3D faces using scanners is more accurate, the entire implementation process is complicated and time-consuming. Therefore, Volker Blanz et al. [**?**] proposed the 3D Morphable Model (3DMM). The 3D morphable model is built based on a 3D face database, using face shape and texture statistics as constraints. The influence of pose and illumination factors on the facial reconstruction process is considered to make the generated 3D face model more accurate. Later, Paysan et al. [**?**] proposed a Basel Face Model (BFM) based on the original 3DMM with improved alignment algorithms for higher shape and texture accuracy. However, most of these methods establish vertex correspondence between the input image and the 3D model through complex fitting processes.

Recent methods bypass problems associated with constructing and fitting 3D deformation models by using a single 2D face image to regress the volume of 3D face geometry for reconstruction. These methods are no longer restricted to the model space but require complex network structures and considerable time to predict voxel data. Recently, some works decompose a given 3D face into identity and expression parts and encode them nonlinearly to achieve 3D face reconstruction with remarkable results. However, these methods do not consider

the interaction between face attributes and only estimate attributes individually during parameter regression. Unlike the above methods, our approach enhances semantically meaningful face attribute representation and directly obtains complete 3D face geometry and its corresponding information by learning the correlation of different 3D face attribute parameters.

III. PROPOSED METHOD

In this section, we first introduce a 3D face model with latent representations and design our approach based on it. Then we propose a transformer-based joint learning pipeline for encoders and decoders. Finally, we provide specific implementation details of this face attribute branching method, including network structure, training data, and training procedure.

A. A Composite 3D Face Shape Model

The 3D Morphable Model is one of the most successful methods for describing 3D face reconstruction. In this work, we adopt a common practice in 3DMM [**?**], representing a 3D human face as a combination of shape and expression. Each 3D face shape is represented as the concatenation of its vertex coordinates:

$$S = [x_1, y_1, z_1, x_2, y_2, z_2, \cdots, x_n, y_n, z_n]^T$$

where n is the number of vertices in the 3D face point cloud, and T denotes transpose. $S_i = (x_i, y_i, z_i)$ represents the (x, y, z) coordinates in the Cartesian coordinate system. In this paper, we use a 3D Morphable Model (3DMM) [**?**] to recover the 3D geometry of a human face from a single image.

We use the 3DMM PCA model to represent the face geometry S_{Model} as:

$$S_{Model} = \bar{S} + \Delta S_{id} + \Delta E_{exp} = \bar{S} + A_{id}\alpha_{id} + B_{exp}\beta_{exp}$$

where $\bar{S} \in \mathbb{R}^{3n}$ is the mean geometry, ΔS_{id} is the identity-sensitive difference between $\bar{S}$ and S_{Model}, and ΔE_{exp} denotes the expression-sensitive difference. A_{id} and α_{id} are the identity basis and identity parameters of the face, while B_{exp} and β_{exp} are the expression basis and expression parameters. $\bar{S}$ and A_{id} are learned from the Basel Face Model [**?**], and B_{exp} is obtained from FaceWarehouse [**?**].

In the 3D face fitting process, we use weak perspective projection to project the 3D face onto the 2D face plane:

$$C(p) = f \cdot Pr \cdot R \cdot S + t$$

where C is the geometry projected in image coordinates, f is a scale factor, Pr is an orthogonal projection matrix, R is a rotation matrix consisting of 9 parameters, and t is a translation vector. We can thus transform the 3D face reconstruction problem into a face parameter regression problem. We have 62 parameters to regress for the 3D face model, where the pose parameter $v_{pose} = [f, R, t]$, so the complete set of model parameters is $p = [v_{pose}, \alpha_{id}, \beta_{exp}]^T$.

B. Transformer Module

Previous approaches [**?**] used disentangled representation to extract individual face attributes from face images. However, these approaches cannot simultaneously model shape, expression, and texture without decomposing the underlying representation into relevant factors such as identity and expression while preserving identity information. Therefore, we use an encoder to extract three different sets of features: identity, expression, and pose. Inspired by [**?**], the Transformer architecture can extract global information from input features and implement information exchange within each entry, whereas the small receptive field of convolutional neural networks can only mine local information. To achieve an effective combination of global and local information, we introduce the Transformer module inside the encoder and decoder. This module can obtain global representation from shallow layers and extract geometric structure information of the face. Additionally, it can capture more similarity of face attributes between shallow and deep representations and extract deep facial semantic information.

As shown in [Figure 2: see original paper], we introduce the Transformer module inside the encoder and decoder, which plays a vital role in the 3D face space. Semantic features in the image are extracted layer by layer inside the encoder in a structured manner through multi-layer convolutional connections with Transformer blocks. Specifically, the encoder module consists of four layers and a Transformer structure. First, the input face image is convolved by a 3×3 Conv layer to obtain initial face features. Then the encoder network extracts these initial face features layer by layer to gradually increase the number of feature channels. Finally, the Transformer block maps features to a high-dimensional space, where the Transformer structure encodes the face as a series of $2048 \times 7 \times 7$ feature sequences.

Additionally, while the traditional Transformer structure uses a self-attention mechanism to directly compute attention weights at each position through certain operations and then compute the implied vector representation of the entire sequence as a weighted sum, this self-attention mechanism has the drawback of overly focusing on a single attribute position when encoding information at the current position. Therefore, we use a multi-head attention mechanism to learn different attribute behaviors of faces and combine these behaviors with different face attributes as knowledge. The aim is to capture dependencies between individual attributes of the face within a sequence and improve subspace representation.

Moreover, to preserve the local features of the face image and not destroy its spatial structure, in the decoder we do not encode the image with position vectors as in the TRT-ViT [**?**] approach. Instead, we convolve the passed face feature sequence by 1×1 Conv and then input the obtained feature sequence into the Transformer structure. Subsequently, the Transformer module transforms the input face attribute features dimensionally, examines the distribution of each face attribute from the feature space, and calculates correlations between face attribute feature points.

C. Attribute Decomposition Representation

Single-view 3D face reconstruction aims to obtain estimates of shape α_{id}, expression β_{exp}, and pose parameters v_{pose} given an input image I. Various attribute sources in the face may lead to variability in facial reconstruction. For example, expressions may cause distortion of geometric contours. Therefore, our goal is to obtain the representational values of each latent attribute from the face image through a function F_i given input image I:

$$(\bar{E}_{id}, \bar{E}_{exp}, \bar{E}_{pose}) = F_i(I; \alpha_{id}, \beta_{exp}, v_{pose})$$

where α_{id}, β_{exp}, and v_{pose} are the identity, expression, and pose parameters involved in F_i. Typically, the latent face attributes $\bar{E}_{id}$, $\bar{E}_{exp}$, and $\bar{E}_{pose}$ are represented in much lower dimensions than the input 2D face image I and output 3D face shape S.

Alternatively, previous approaches extracted shape, expression, and pose attribute information independently given input image I:

$$\bar{E}_{id} = F_{id}(I; \alpha_{id}), \quad \bar{E}_{exp} = F_{exp}(I; \alpha_{exp}), \quad \bar{E}_{pose} = F_{pose}(I; \alpha_{pose})$$

However, this simple feature extraction does not consider correlations between face attributes but only decomposes low-dimensional 3D faces across attribute variables.

We designed a Disentangled Representation Transformer Network (DRTN) to solve this problem. The network structure is shown in [Figure 3: see original paper], where dependencies between face attributes are learned through regression from identity, expression, and pose branches. The specific learning of face attributes for each branch is described below.

Identity Branch: One branch of the identity component aims to enhance learning of expression and pose attributes by preserving the overall facial geometry structure while keeping identity unchanged. In the identity branch, we explicitly model individual face attribute dependencies, where joint expression and pose attributes are decomposed under the condition that the identity attribute $\bar{E}_{id} = F_{id}(I; \alpha_{id})$ remains consistent:

$$\bar{E}_{id,exp} = F_{id,exp}(\beta_{exp}; I, \bar{E}_{id})$$
$$\bar{E}_{id,exp,pose} = F_{id,exp,pose}(v_{pose}; I, \bar{E}_{id}, \bar{E}_{id,exp})$$
$$\bar{E}_{id,pose} = F_{id,pose}(v_{pose}; I, \bar{E}_{id})$$
$$\bar{E}_{id,pose,exp} = F_{id,pose,exp}(\beta_{exp}; I, \bar{E}_{id}, \bar{E}_{id,pose})$$

We formulate the learning process of these parameters with three autoencoders E_{id}, E_{exp}, and E_{pose}. $\bar{E}_{id}$ is the identity attribute information learned from input image I after passing through E_{id}. $\bar{E}_{id,exp}$ is the expression information obtained by E_{exp} encoder learning based on identity attribute $\bar{E}_{id}$. $\bar{E}_{id,exp,pose}$ is the pose information learned by the E_{pose} encoder based on $\bar{E}_{id,exp}$. Similarly, $\bar{E}_{id,pose}$ is the pose information obtained by E_{pose} encoder learning based on identity property $\bar{E}_{id}$. $\bar{E}_{id,pose,exp}$ is the expression information obtained by E_{exp} encoder learning based on $\bar{E}_{id,pose}$. $F_i(\cdot)$ represents the learnable encoder among autoencoders that have undergone different processing orders.

Although this method can effectively solve the problem of non-interactive information between face attributes, parameter estimation using this sequential decomposition approach is somewhat scattered. To address this, we use a coupled variable method to fuse expression and pose attribute information under identity invariance, with the fused attribute representation value defined as:

$$\bar{T}(\alpha_{id}, \beta_{exp}, v_{pose}) = \bar{E}_{id} \otimes \bar{E}_{id,exp} \otimes \bar{E}_{id,exp,pose}$$
$$\bar{T}(\alpha_{id}, v_{pose}, \beta_{exp}) = \bar{E}_{id} \otimes \bar{E}_{id,pose} \otimes \bar{E}_{id,pose,exp}$$

where $\bar{T}(\alpha_{id}, \beta_{exp}, v_{pose})$ and $\bar{T}(\alpha_{id}, v_{pose}, \beta_{exp})$ represent the face expression and pose features obtained after facial attributes pass through different encoder networks based on identity consistency, and $\otimes$ denotes the element-wise Hadamard product.

Expression Branch: Correspondingly, the expression branch couples identity and pose attributes to refine facial detail reconstruction while preserving expression attribute consistency $\bar{E}_{exp} = F_{exp}(I; \beta_{exp})$. Its joint attribute decomposition is expressed as:

$$\bar{E}_{exp,id} = F_{exp,id}(\alpha_{id}; I, \bar{E}_{exp})$$
$$\bar{E}_{exp,id,pose} = F_{exp,id,pose}(v_{pose}; I, \bar{E}_{exp}, \bar{E}_{exp,id})$$
$$\bar{E}_{exp,pose} = F_{exp,pose}(v_{pose}; I, \bar{E}_{exp})$$
$$\bar{E}_{exp,pose,id} = F_{exp,pose,id}(\alpha_{id}; I, \bar{E}_{exp}, \bar{E}_{exp,pose})$$

where $\bar{E}_{exp}$ is the expression attribute information obtained from input image I after E_{exp} encoder learning. $\bar{E}_{exp,id}$ is the identity information obtained by E_{id} encoder learning based on expression attribute $\bar{E}_{exp}$. $\bar{E}_{exp,id,pose}$ is the pose information obtained by E_{id} encoder learning based on $\bar{E}_{exp,id,pose}$. Similarly, $\bar{E}_{exp,pose}$ is the pose information obtained by E_{pose} encoder learning based on expression property $\bar{E}_{exp}$. $\bar{E}_{exp,pose,id}$ is the identity information obtained by E_{id} encoder learning based on $\bar{E}_{exp,pose}$. The representation values of identity and pose attributes under constant expression are:

$$\bar{T}(\beta_{exp}, \alpha_{id}, v_{pose}) = \bar{E}_{exp} \otimes \bar{E}_{exp,id} \otimes \bar{E}_{exp,id,pose}$$
$$\bar{T}(\beta_{exp}, v_{pose}, \alpha_{id}) = \bar{E}_{exp} \otimes \bar{E}_{exp,pose} \otimes \bar{E}_{exp,pose,id}$$

where $\bar{T}(\beta_{exp}, \alpha_{id}, v_{pose})$ and $\bar{T}(\beta_{exp}, v_{pose}, \alpha_{id})$ represent the face pose and identity features obtained after facial attributes pass through different encoder networks based on expression consistency.

Pose Branch: The pose branch aims to improve 3D face landmark alignment by coupling identity and pose attributes while preserving pose attribute consistency $\bar{E}_{pose} = F_{pose}(I; v_{pose})$. Its joint attribute decomposition is expressed as:

$$\bar{E}_{pose,id} = F_{pose,id}(\alpha_{id}; I, \bar{E}_{pose})$$
$$\bar{E}_{pose,id,exp} = F_{pose,id,exp}(\beta_{exp}; I, \bar{E}_{pose}, \bar{E}_{pose,id})$$
$$\bar{E}_{pose,exp} = F_{pose,exp}(\beta_{exp}; I, \bar{E}_{pose})$$
$$\bar{E}_{pose,exp,id} = F_{pose,exp,id}(\alpha_{id}; I, \bar{E}_{pose}, \bar{E}_{pose,exp})$$

where $\bar{E}_{pose}$ is the pose attribute information obtained from input image I after E_{pose} learning. $\bar{E}_{pose,id}$ and $\bar{E}_{pose,exp}$ are the identity and expression information obtained by E_{id} and E_{exp} encoders based on pose attribute $\bar{E}_{pose}$. $\bar{E}_{pose,id,exp}$ is the expression information obtained by E_{exp} encoder learning based on $\bar{E}_{pose,id}$. Similarly, $\bar{E}_{pose,exp,id}$ is the identity information obtained by E_{id} encoder learning based on $\bar{E}_{pose,exp}$. The standard attribute representation values of identity and expression under constant pose are:

$$\bar{T}(v_{pose}, \alpha_{id}, \beta_{exp}) = \bar{E}_{pose} \otimes \bar{E}_{pose,id} \otimes \bar{E}_{pose,id,exp}$$
$$\bar{T}(v_{pose}, \beta_{exp}, \alpha_{id}) = \bar{E}_{pose} \otimes \bar{E}_{pose,exp} \otimes \bar{E}_{pose,exp,id}$$

where $\bar{T}(v_{pose}, \alpha_{id}, \beta_{exp})$ and $\bar{T}(v_{pose}, \beta_{exp}, \alpha_{id})$ represent the face expression and identity features obtained after facial attributes pass through different encoder networks based on pose consistency.

Fusion Module: To reconstruct a more realistic and complete 3D human face, we extract face information related to each attribute from the identity, expression, and pose branching networks, then use the fusion module to merge these attributes to complete accurate 3D face reconstruction. This ensures the validity of face attribute decomposition. Additionally, face images passing through low-level networks typically have higher resolution and contain more detailed information. Therefore, we introduce shallow face attribute information from each branch in the fusion module to better refine facial details. The face attribute features of the fusion module are represented as:

$$G_i = T_i(R(\bar{T}(\alpha_{id}, \beta_{exp}, v_{pose}), \bar{T}(\alpha_{id}, v_{pose}, \beta_{exp}), \bar{T}(\beta_{exp}, \alpha_{id}, v_{pose}), \bar{T}(\beta_{exp}, v_{pose}, \alpha_{id}), \bar{T}(v_{pose}, \alpha_{id}, \beta_{exp}), \bar{T}(v_p$$

where $T_i(\cdot)$ denotes feature information after fusion of face attributes, $R(\cdot)$ denotes fusion of identity, expression, and pose attributes, and $G_i(\cdot)$ is the final feature output after fusion.

D. Objective Loss Function

In the face attribute branching network, we introduce four learning objectives for model training. For learning face attribute parameters, we constrain from three parts: identity, expression, and pose. To improve the network' s effective learning of identity attributes, we optimize parameters α_{id} by minimizing the vertex distance between predicted face geometry and ground truth 3D face:

$$L_{id} = \|S_{model} - \bar{S}_{model}\| = \|\Delta S_{id} - \Delta \bar{S}_{id}\| = \|(A_{id}\alpha_{id} - A_{id}\bar{\alpha}_{id})\|_2$$

where α_{id} denotes the predicted face identity parameter, $\bar{\alpha}_{id}$ is the ground truth parameter, and A_{id} is the identity basis of the 3D deformation model PCA. Additionally, different α_{id} dimensions and singular values affect face geometry differently. Therefore, the L_{id} constraint on identity information can reduce the influence of essential dimensions, especially those with large singular values.

Similarly, to refine facial expression details, we use expression consistency loss to enhance expression preservation:

$$L_{exp} = \|\Delta E_{exp} - \Delta \bar{E}_{exp}\| = \|(B_{exp}\beta_{exp} - B_{exp}\bar{\beta}_{exp})\|$$

where β_{exp} denotes the predicted face expression parameter, $\bar{\beta}_{exp}$ is the expression ground truth parameter, and B_{exp} is the expression basis of the 3D deformation model PCA.

Since face pose has limited degrees of freedom, to simplify computation we constrain pose estimation through facial landmark loss:

$$L_p = \|(f \cdot Pr \cdot R \cdot \bar{S}_{model} + t) - (\bar{f} \cdot Pr \cdot \bar{R} \cdot \bar{S}_{model} + \bar{t})\|_2$$

where $v_{pose} = [f, R, t]$ is the predicted pose parameter and $\bar{v}_{pose} = [\bar{f}, \bar{R}, \bar{t}]$ is the ground truth pose parameter. This further improves face landmark alignment accuracy under significant pose conditions by constraining pose parameters.

Although L_{id}, L_{exp}, and L_{pose} loss functions provide strong constraints on identity, expression, and pose attributes across the three branches, reconstructed 3D faces still lack geometric contour constraints. Therefore, we use sparse 2D face landmark constraints to improve reconstructed geometric contour information:

$$L_{68} = \|(f \cdot Pr \cdot R \cdot S_{68} + t) - (\bar{f} \cdot Pr \cdot \bar{R} \cdot \bar{S}_{68} + \bar{t})\|_2$$

The final loss function L is defined as:

$$L = \lambda_{id} L_{id} + \lambda_{exp} L_{exp} + \lambda_p L_p + \lambda_{68} L_{68}$$

where λ_{id}, λ_{exp}, λ_{pose}, and λ_{68} are weights to balance these constraints.

IV. EXPERIMENTS

In this section, we perform several experiments and ablation studies on DRTN using different settings on three extensively evaluated datasets—300W-LP [**?**], AFLW2000-3D [**?**], and AFLW [**?**]—to demonstrate the method' s effectiveness in 3D face reconstruction and dense face alignment. Additionally, we evaluate the generalization performance of DRTN on the HELEN [**?**], LS3D-W [**?**], and CelebA [**?**] datasets.

A. Implementation Details

We use the PyTorch [**?**] deep learning framework to train the DRTN model. For training, face regions are cropped according to ground truth 3D facial landmarks and scaled to 256×256 as network input. Throughout branching network training, increasing network layers and parameters slows neural network training and occasionally causes overfitting to the training set, potentially preventing deep cascade regression networks from learning useful information from overfitted samples. To better balance learning weights of each branch parameter, we set loss weights of $\lambda_{id} = 1.0 \times 10^{-3}$, $\lambda_{exp} = 1.0 \times 10^{-4}$, and $\lambda_{pose} = 1.0 \times 10^{-4}$ after extensive experimental tuning.

All training and testing experiments were conducted on a PC with NVIDIA GeForce GPU and CUDA 11.2. The SGD solver was used with a minimum batch size of 128 and initial learning rate of 0.03. Our training set contained 687,854 face images, including 122,450 natural face images and 565,404 synthetic face images. Natural face images are from the 300W-LP [**?**] dataset, extended using various data augmentation algorithms. We performed a total of 70 training

batches, reducing the learning rate to 0.02, 0.004, and 0.0008 after 30, 40, and 60 batches, respectively.

B. Datasets

The proposed DRTN is trained on 300W-LP [?] and evaluated on five popular datasets: AFLW [?], AFLW2000-3D [?], LS3D-W [?], and CelebA [?].

300W-LP: The 300W-LP dataset [?] is used to train our model, extended from 300W by standardizing multiple alignment datasets with 68 landmarks, including AFW [?], LFPW [?], HELEN [?], IBUG [?], and XM2VTS [?].

AFLW2000-3D: AFLW2000-3D [?] is constructed to evaluate 3D face alignment on challenging unconstrained images. This dataset contains the first 2,000 images from AFLW and expands its annotations with fitted 3DMM parameters and 68 3D landmarks. We use this dataset to evaluate our method's performance for face alignment and reconstruction tasks.

AFLW: AFLW [?] is a large-scale face dataset including multiple poses and views, commonly used to evaluate facial landmark detection effectiveness. The dataset has 25,993 face images with 21 landmarks annotated per face, though landmarks are not annotated for invisible regions. The dataset also includes face pose angle annotations obtained from average 3D face reconstruction. Most AFLW images are color images, with a few grayscale images; 59% are female and 41% are male. This dataset is well-suited for multi-angle multi-face detection, landmark localization, and head pose estimation, making it essential for face landmark alignment research.

LS3D-W: LS3D-W [?] is a large-scale face alignment annotation dataset created by the Computer Vision Laboratory at the University of Nottingham. Face images are from AFLW [?], 300-VW [?], 300-W [?], and FDDB [?]. Each face image contains 68 annotated landmarks, totaling approximately 230,000 accurately labeled face images.

CelebA: CelebA [?] is a large-scale face attribute dataset containing over 200K images, each with 40 attribute annotations. Images cover extensive pose variations and complex backgrounds. The CelebA dataset includes 10,177 identities and 202,599 face images.

C. Evaluation

For face alignment and reconstruction, we quantitatively evaluate performance using NME2d and NME3d metrics:

$$NME_{2d} = \frac{\sum_{i=1}^{N_k} \|V_i^{2d} - \bar{V}_i^{2d}\|_2}{\sum_{i=1}^{N_k} D_j}$$

$$NME_{3d} = \frac{\sum_{i=1}^{N_k} \|V_i^{3d} - \bar{V}_i^{3d}\|_2}{\sum_{i=1}^{N_k} S_j}$$

where V_i^{2d} and V_i^{3d} denote estimated 2D landmarks and 3D vertices, and $\bar{V}_i^{2d}$ and $\bar{V}_i^{3d}$ are ground truth landmarks and vertices for N_k test images. D_j and S_j are the diagonal sizes of the face region in image space and 3D coordinate space, respectively. NME_{2d} evaluates normalized 2D facial landmark prediction error, while NME_{3d} evaluates normalized 3D face geometry estimation accuracy. Due to ambiguity in weak perspective projection reconstruction results, different methods have some ambiguity in the z-axis direction. We use rigid translation along the z-axis to align each result with ground truth.

D. Analysis of 3D Face Alignment Results

We quantitatively evaluate face landmark alignment performance using normalized mean error $NME_{2D}(\%)$ on AFLW2000-3D [**?**] and AFLW [**?**] datasets. We divide the test set into three subsets based on absolute yaw angle: $[0^\circ, 30^\circ]$, $[30^\circ, 60^\circ]$, and $[60^\circ, 90^\circ]$. Our DRTN was compared experimentally with current state-of-the-art methods [**?**, **?**, **?**, **?**, **?**]. Results are shown in , with best results in each category highlighted in bold (lower values are better). [Figure 4: see original paper] shows corresponding CED curves, comparing our DRTN only with methods having available code from Table 1. Other methods are not comparable primarily due to lack of open-source code.

Compared to benchmark methods [**?**, **?**, **?**, **?**, **?**, **?**], the DRTN method achieves lower normalized mean error on AFLW [**?**] and AFLW-2000 [**?**] datasets. Experimental results show that DRTN significantly improves 3D face landmark alignment accuracy across the full pose range, with robust landmark alignment even under large pose conditions.

[Figure 5: see original paper] shows our method' s 3D face landmark alignment results on AFLW [**?**] and AFLW2000-3D [**?**] datasets. The advantage of using 3DMM over other geometry representations is that normal mapping can associate semantic facial landmarks with corresponding points in reconstructed geometry. Visualization results demonstrate that DRTN significantly outperforms 3DDFA [**?**], DAMDNet [**?**], GSRN [**?**], MARN [**?**], and MFRRN [**?**] methods for 3D face landmark alignment under large pose, extreme expression, and occlusion conditions, particularly for eyes, mouth, and face contours. Qualitative results show that our DRTN method significantly improves network learning of face attribute features in complex environments, thereby enhancing face landmark alignment accuracy.

E. Analysis of 3D Face Reconstruction Results

The AFLW dataset is unsuitable for evaluating 3D face reconstruction because recovering 3D faces from annotated visible landmarks usually leads to ambiguity problems. To validate our model' s effectiveness in 3D face reconstruction, we compare 3D normalized mean error $NME_{3d}(\%)$ on the AFLW2000-3D dataset, as shown in (first-best results in each category highlighted in bold; lower is better). Table 2 results show that our DRTN method outperforms state-of-the-art methods [**?**, **?**, **?**, **?**, **?**] for face reconstruction in both medium and large poses. Compared to recent MARN [**?**], DRTN reduces $NME_{3d}(\%)$ by 3.2%, 3.9%,

and 4.9% at offset angles $[0°, 30°]$, $[30°, 60°]$, and $[60°, 90°]$, respectively. Our method significantly improves prediction accuracy at large offset angle $[60°, 90°]$, indicating the model can accurately reconstruct 3D faces in small, medium, and large poses. These results fully validate the robustness of our DRTN algorithm for reconstruction tasks under large pose conditions.

Visualization results are shown in [Figure 8: see original paper]. Our DRTN method reconstructs 3D faces with clear texture and natural expressions under large poses, extreme expressions, and partial occlusion. As shown in [Figure 9: see original paper], DRTN achieves more geometrically accurate reconstructed facial geometry compared to current state-of-the-art methods, particularly regarding local details of eyes, mouth, and wrinkles. Methods such as 3DDFA [**?**], DAMDNet [**?**], MARN [**?**], and MFIRRN [**?**], due to lack of face attribute correlation learning, cannot reconstruct faces with fine expression details. Therefore, our DRTN approach offers significant advantages for high-fidelity 3D facial reconstruction under large poses, occlusion, and extreme expressions.

[Figure 6: see original paper] shows the corresponding CED for 3D face reconstruction, intuitively proving that DRTN can achieve accurate 3D face reconstruction.

To further demonstrate our method' s effectiveness on face reconstruction in large poses, we regard the entire AFLW2000-3D as the test set and divide it into three subsets according to absolute yaw angles: $[0°, 30°]$, $[30°, 60°]$, and $[60°, 90°]$ with 1,312, 383, and 305 samples respectively. [Figure 7: see original paper] shows CED curve results from left to right for small pose $[0°, 30°]$, medium pose $[30°, 60°]$, and large pose $[60°, 90°]$. Comparing CED curves, our method demonstrates improvement even under unconstrained large pose conditions. Compared with GSRN [**?**], DRTN reduces $NME_{3d}(\%)$ by 1.0% and 6.6% at medium pose $[30°, 60°]$ and large pose $[60°, 90°]$, respectively, further validating the superior 3D reconstruction capability of our DRTN model under complex occlusions.

F. Ablation Experiments

To verify the effectiveness of each attribute branch module in DRTN, presents ablation experiments on the AFLW-2000 dataset, where normalized mean errors of face landmark alignment and reconstruction are NME_{2D} and NME_{3D}, respectively. Comparing the first four rows of Table 3 shows that networks with face pose, expression, and identity attribute branches achieve lower reconstruction and alignment errors than the baseline model without any face attribute branches. These results demonstrate that disentangled representation of face attributes can improve the model' s attribute learning ability and robustness.

Additionally, rows 2-4 of Table 3 reveal that the identity attribute model outperforms expression and pose attribute models in single-branch face attribute networks because training dataset labels contain more identity attribute parameters than expression and pose parameters. Therefore, when labels contain more annotation information in a single-branch face attribute network, it is

more beneficial for the network to learn attribute parameters. Similarly, rows 5-7 show that network models using two-branch face attributes achieve lower face alignment and reconstruction errors than single-branch face attribute networks. These results further indicate that using disentangled face attribute representation can improve network learning of face attribute features. Finally, error results demonstrate that our DRTN method shows good attribute learning ability in face alignment and reconstruction, primarily because information between face attributes is correlated rather than existing singularly. We can enrich face reconstruction details through attribute correlations.

To further demonstrate the role of face attribute branching networks in face landmark alignment and reconstruction, we visualize results for each branch network in [Figure 10: see original paper]. The reconstructed high-fidelity 3D face and geometry results show that 3D faces reconstructed by our DRTN method have natural expressions, fit original image shapes well, and exhibit clear facial edge geometry and five-feature contours.

We visualize reconstructed 3D faces of each model as heat maps, which show that our method produces more realistic 3D faces with fewer errors. Regarding 3D face landmark alignment, DRTN also achieves higher landmark alignment accuracy than other face attribute branching models in large poses. Other face attribute branching networks lack mutual learning between attributes during parameter learning. Although these networks learn independent face attributes well, reconstruction and alignment accuracy in complex and diverse situations could be improved because 3D faces are not linear models. In contrast, the DRTN model incorporates identity, expression, and pose information, demonstrating good stability in reconstructing unconstrained scenes.

G. Visualization Experiment on LS3D-W

This section compares qualitative results of our DRTN method with state-of-the-art methods on LS3D-W [**?**]. [Figure 11: see original paper] shows 3D face reconstruction and landmark alignment results in complex scenes with insufficient illumination and large face poses. The red dashed box demonstrates that DRTN achieves higher reconstruction accuracy than 3DDFA [**?**], DAMDNet [**?**], MARN [**?**], and MFIRRN [**?**] models in high-fidelity 3D face reconstruction, particularly in facial feature contours, validating that our model better captures local details and geometric contour information. LS3D-W [**?**] is collected in highly unconstrained complex environments with varying illumination and face offset angles. The blue dashed box shows face landmark alignment results, where DRTN maintains good alignment performance even under self-occlusion and low-light conditions. These excellent results verify the powerful 3D reconstruction capability and landmark alignment accuracy of the face attribute disentangled representation method.

H. Qualitative Results on CelebA and Casual Photos

To further validate DRTN' s generalization ability on other datasets, we evaluated it on casual photos and CelebA [**?**]. As shown in [Figure 12: see original

paper], face images in the green dashed box show CelebA [**?**] results, while those in the red dashed box are from random life photos and anime character photos. Our DRTN accomplishes accurate landmark alignment and fine 3D face reconstruction on both datasets, reflecting good generalization ability. Because DRTN enables the network to learn correlations between identity, expression, and pose attributes and fuse underlying information between face attributes, it can reconstruct accurate high-fidelity 3D faces and perform 3D face landmark alignment.

V. CONCLUSION

In this paper, we propose a Disentangled Representation Transformer Network capable of recovering detailed 3D faces in unconstrained environments and performing dense face alignment more accurately. Our DRTN method enhances network learning of latent face attribute information and addresses effects of facial expression, head pose, and partial occlusion on reconstruction and landmark alignment. Quantitative results show that DRTN is more accurate than state-of-the-art dense face alignment and 3D reconstruction methods. Additionally, extensive qualitative experiments demonstrate that DRTN can successfully reconstruct high-fidelity 3D faces from 2D images with rich details and strong generalization ability. In future work, we will further investigate the proposed method for 3D face reconstruction in videos and cartoon character reconstruction to evaluate its generalization capability more comprehensively.

REFERENCES

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

[2] Hu, J., Lu, J., Tan, Y. P. (2014). Discriminative deep metric learning for face verification in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1875-1882).

[3] Bettadapura, V. (2012). Face expression recognition and analysis: the state of the art. *arXiv preprint arXiv:1203.6722*.

[4] Cao, C., Wu, H., Weng, Y., Shao, T., Zhou, K. (2016). Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics*, 35(4).

[5] Amberg, B., Romdhani, S., Vetter, T. (2007, June). Optimal step non-rigid ICP algorithms for surface registration. In *2007 IEEE conference on computer vision and pattern recognition* (pp. 1-8). IEEE.

[6] Kemelmacher-Shlizerman, I., Basri, R. (2010). 3D face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence*, 33(2), 394-405.

[7] Li, C., Zhou, K., Lin, S. (2014, September). Intrinsic face image decomposition with human face priors. In *European conference on computer vision* (pp. 218-233). Springer, Cham.

[8] Cao, X., Chen, Z., Chen, A., Chen, X., Li, S., Yu, J. (2018). Sparse photometric 3D face reconstruction guided by morphable models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4635-4644).

[9] Tuan Tran, A., Hassner, T., Masi, I., Medioni, G. (2017). Regressing robust and discriminative 3D morphable models with a very deep neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5163-5172).

[10] Jourabloo, A., Liu, X. (2016). Large-pose face alignment via CNN-based dense 3D model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4188-4196).

[11] Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 534-551).

[12] Jackson, A. S., Bulat, A., Argyriou, V., Tzimiropoulos, G. (2017). Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE international conference on computer vision* (pp.1031-1039).

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

[14] Li, X., Wu, S. (2021, January). Multi-attribute regression network for face reconstruction. In *2020 25th International Conference on Pattern Recognition (ICPR)* (pp. 7226-7233). IEEE.

[15] Cootes, T. F., Taylor, C. J., Cooper, D. H., Graham, J. (1995). Active shape models-their training and application. *Computer vision and image understanding*, 61(1), 38-59.

[16] Cootes, T. F., Edwards, G. J., Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6), 681-685.

[17] Cootes, T. F., Ionita, M. C., Lindner, C., Sauer, P. (2012, October). Robust and accurate shape model fitting using random forest regression voting.

In *European conference on computer vision* (pp. 278-291). Springer, Berlin, Heidelberg.

[18] Jourabloo, A., Liu, X. (2015). Pose-invariant 3D face alignment. In *Proceedings of the IEEE international conference on computer vision* (pp. 3694-3702).

[19] Bulat, A., Tzimiropoulos, G. (2016, October). Two-stage convolutional part heatmap regression for the 1st 3d face alignment in the wild (3dfaw) challenge. In *European Conference on Computer Vision* (pp. 616-624). Springer, Cham.

[20] Jeni, L. A., Cohn, J. F., Kanade, T. (2015, May). Dense 3D face alignment from 2D videos in real-time. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (Vol. 1, pp. 1-8). IEEE.

[21] Ning, X., Duan, P., Li, W., Zhang, S. (2020). Real-time 3D face alignment using an encoder-decoder network with an efficient deconvolution layer. *IEEE Signal Processing Letters*, 27, 1944-1948.

[22] Zhu, X., Lei, Z., Liu, X., Shi, H., Li, S. Z. (2016). Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 146-155).

[23] Liu, F., Zeng, D., Zhao, Q., Liu, X. (2016, October). Joint face alignment and 3D face reconstruction. In *European Conference on Computer Vision* (pp. 545-560). Springer, Cham.

[24] Jiang, L., Wu, X. J., Kittler, J. (2019). Dual attention MobDenseNet (DAMDNet) for robust 3D face alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops* (pp. 0-0).

[25] Ye, H., Lv, L., Liu, Y., Zhou, Y. (2016). Evaluation of the Accuracy, Reliability, and Reproducibility of Two Different 3D Face-Scanning Systems. *The International Journal of Prosthodontics*, 29(3), 213-218.

[26] De Wansa Wickramaratne, V. K., Ryazanov, V. V., Vinogradov, A. P. (2009). Analysis of a 3D face-scanning system by active triangulation. *Pattern Recognition and Image Analysis*, 19(1), 78-83.

[27] Blanz, V., Vetter, T. (1999, July). A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques* (pp. 187-194).

[28] Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T. (2009, September). A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance* (pp. 296-301). IEEE.

[29] Richardson, E., Sela, M., Or-El, R., Kimmel, R. (2017). Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1259-1268).

[30] Dou, P., Shah, S. K., Kakadiaris, I. A. (2017). End-to-end 3D face reconstruction with deep neural networks. In *proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5908-5917).

[31] Tran, L., Liu, X. (2018). Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7346-7355).

[32] Lee, G. H., Lee, S. W. (2020). Uncertainty-aware mesh decoder for high fidelity 3D face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6100-6109).

[33] Wu, S., Rupprecht, C., Vedaldi, A. (2020). Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1-10).

[34] Browatzki, B., Wallraven, C. (2020). 3fabrec: Fast few-shot face alignment by reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6110-6120).

[35] Feng, Y., Wu, F., Shao, X., Wang, Y., Zhou, X. (2018). Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 534-551).

[36] Cao, C., Weng, Y., Zhou, S., Tong, Y., Zhou, K. (2013). Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3), 413-425.

[37] Xia, X., Li, J., Wu, J., Wang, X., Wang, M., Xiao, X., ···Wang, R. (2022). TRT-ViT: TensorRT-oriented Vision Transformer. *arXiv preprint arXiv:2205.09579*.

[38] Koestinger, M., Wohlhart, P., Roth, P. M., Bischof, H. (2011, November). Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 2144-2151). IEEE.

[39] Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q. (2013). Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 386-391).

[40] Liu, Z., Luo, P., Wang, X., Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision* (pp. 3730-3738).

[41] Bulat, A., Tzimiropoulos, G. (2017). How far are we from solving the 2d 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Pro-*

ceedings of the IEEE International Conference on Computer Vision (pp. 1021-1030).

[42] Paszke, A., Gross, S., Chintala, S., Chanan, G. (2017). Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 6(3), 67.

[43] Zhu, X., Ramanan, D. (2012, June). Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 2879-2886). IEEE.

[44] Belhumeur, P. N., Jacobs, D. W., Kriegman, D. J., Kumar, N. (2013). Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12), 2930-2940.

[45] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 397-403).

[46] 300 Faces in-the-Wild Challenge, accessed on Jul.2013. [Online]. Available: http://ibug.doc.ic.ac.uk/resources/300-W/.

[47] Shen, J., Zafeiriou, S., Chrysos, G. G., Kossaifi, J., Tzimiropoulos, G., Pantic, M. (2015). The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 50-58).

[48] Jain, V., Learned-Miller, E. (2010). Fddb: A benchmark for face detection in unconstrained settings (Vol. 2, No. 6). UMass Amherst technical report.

[49] Yu, X., Huang, J., Zhang, S., Metaxas, D. N. (2015). Face landmark fitting via optimized part mixtures and cascaded deformable model. *IEEE transactions on pattern analysis and machine intelligence*, 38(11).

[50] Burgos-Artizzu, X. P., Perona, P., Dollár, P. (2013). Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision* (pp. 1513-1520).

[51] Cao, X., Wei, Y., Wen, F., Sun, J. (2014). Face alignment by explicit shape regression. *International journal of computer vision*, 107(2), 177-190.

[52] Yan, J., Lei, Z., Yi, D., Li, S. (2013). Learn to combine multiple hypotheses for accurate face alignment. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 392-396).

[53] Liu, Y., Jourabloo, A., Ren, W., Liu, X. (2017). Dense face alignment. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 1619-1628).

[54] Yu, R., Saito, S., Li, H., Ceylan, D., Li, H. (2017). Learning dense facial correspondences in unconstrained images. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 4723-4732).

[55] Wang, X., Li, X., Wu, S. (2021, June). Graph structure reasoning network for face alignment and reconstruction. In *International Conference on Multimedia Modeling* (pp. 493-505). Springer, Cham.

[56] Li, L., Li, X., Wu, K., Lin, K., Wu, S. (2021, June). Multi-granularity feature interaction and relation reasoning for 3d dense alignment and face reconstruction. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4265-4269). IEEE.

Note: Figure translations are in progress. See original paper for figures.