

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

Supplementary Material

Xiangzheng Li¹, Xiaofei Li², Jia Deng¹ and Xiangzheng Li^{1*}

^{1*}School of Mathematics and Computer Science, Ningxia Normal University, Xueyuan Road, Guyuan, 756000, Ningxia, China.

^{1,2}School of Mathematics and Computer Science, Ningxia Normal University, Xueyuan Road, Guyuan, 756000, Ningxia, China.

*Corresponding author(s). E-mail(s): 82021011@nxnu.edu.cn;

Contributing authors: lxz_nxu@163.com; 2080755443@qq.com;
d792254122@163.com;

1 3D Face Reconstruction Process

The process of 3D face reconstruction involves several steps. First, the facial region is extracted from the input image. Then, a pre-trained model is used for forward propagation to obtain the model's output parameters, including facial shape parameters α_{id} , expression parameters β_{exp} , and pose parameter v_{pose} . These parameters are combined with a 3DMM to generate the corresponding face model. During the forward propagation process, the three-dimensional vertex coordinates of the face are calculated using the model's output parameters and the facial bounding box. Moreover, the original image is sampled at these computed vertex coordinates to obtain the corresponding color information. This process includes mapping the texture information of the 2D facial image onto the 3D face model, effectively assigning colors to each vertex. Next, the 3D face with texture is rendered using the vertex coordinates and texture mapping. By computing the color values for each vertex and employing the Z-buffer algorithm, the color values are filled into the pixels of the triangle mesh. This step involves depth sorting based on the 3D model's depth information to

ensure that occluded pixels do not overwrite foreground pixels. Consequently, the textured 3D facial model is realistically rendered, and due to capturing the color and details from the original 2D facial image, the reconstructed 3D face appears more natural.

2 Experiments

2.1 Implementation Details

We use the Pytorch [6] deep learning framework to train the DRTN model. The face regions are cropped according to the ground truth 3D facial landmarks to train our framework and then scaled to 256×256 as the input to our network. Training the branching network reveals that as the number of layers and parameters increases, the neural network training slows and occasionally overfits the training set. This results in a deep cascade regression network that may not learn information from the overfitted samples. Hence, the learning weights of each branch parameter and the loss weights are heuristically set to $\lambda_{id} = 1.0e^{-3}$, $\lambda_{exp} = 1.0e^{-4}$ and $\lambda_{pose} = 1.0e^{-4}$ after several experiments.

Both training and testing experiments were conducted on a PC with an NVIDIA GeForce and CUDA 11.2. The SGD solver involved a minimum batch size and initial learning rate of 128 and 0.03, respectively. Besides, 687,854 face images were available for training, involving 122,450 natural face images and 565,404 synthetic face images. The authentic face images are from the 300W-LP[3] dataset, which is extended using various data enhancement algorithms. We performed 70 batches of training, and after 30, 40, and 60 batches, we reduced the learning rate to 0.02, 0.004, and 0.0008, respectively.

2.2 Evaluation

The face alignment and reconstruction performance of each method is quantitatively evaluated using the NME_{2d} and NME_{3d} metrics:

$$NME_{2d} = \frac{1}{N_k} \sum_j^{N_k} \left[\frac{1}{D_j} \sum_i^N (\|V_i^{2d} - \bar{V}_i^{2d}\|_2) \right] \quad (1)$$

$$NME_{3d} = \frac{1}{N_k} \sum_j^{N_k} \left[\frac{1}{S_j} \sum_i^N (\|V_i^{3d} - \bar{V}_i^{3d}\|_2) \right] \quad (2)$$

where V_i^{2d} and $\bar{V}_i^{2d}$ denote the estimated 2D landmarks and the ground truth landmarks, respectively. The ground truth 3D vertices $\bar{V}_i^{3d}$ and the estimated vertices V_i^{3d} for N_k test images are known. D_j and S_j are the diagonal sizes of the face region in the image and 3D coordinate space, respectively. NME_{2d} evaluates the normalized 2D facial landmarks prediction error, and NME_{3d} evaluates the normalized 3D face geometry estimation accuracy. Due to the ambiguity of the weak perspective projection model, the reconstruction results of different methods present some ambiguity in the z-axis direction. Therefore,

Table 1 Normalized Mean Error (*NME*) on AFLW and AFLW2000-3D datasets for network branching structures trained with different losses. Our full hybrid loss function yields significantly higher accuracy than other baselines on both datasets.

$\mathcal{L}_{id}$	Losses			AFLW	AFLW-2000	
	$\mathcal{L}_{exp}$	$\mathcal{L}_{pose}$	$\mathcal{L}_{68}$	NME_{2d}	NME_{2d}	NME_{3d}
		✓		6.399 ± 0.827	4.494 ± 0.893	6.544 ± 1.278
	✓			5.848 ± 0.768	4.276 ± 0.855	6.405 ± 1.166
✓				5.574 ± 0.696	4.038 ± 0.727	5.969 ± 1.182
	✓	✓		5.318 ± 0.616	3.876 ± 0.719	5.883 ± 1.137
✓		✓		5.158 ± 0.632	3.749 ± 0.784	5.722 ± 1.143
✓	✓			5.102 ± 0.594	3.715 ± 0.724	5.694 ± 1.146
✓	✓	✓		4.984 ± 0.581	3.691 ± 0.723	5.681 ± 1.149
✓	✓	✓	✓	4.802 ± 0.536	3.630 ± 0.644	5.669 ± 1.070

we use rigid translation along the z-axis to align each result with the ground truth.

3 More Results and Analysis

3.1 Validating The Hybrid Loss

We conducted specific ablation experiments on the loss functions to investigate the effectiveness of adding a hybrid loss function, as shown in **Table 1**. In the fusion network, we individually applied the $\mathcal{L}_{pose}$, $\mathcal{L}_{exp}$, and $\mathcal{L}_{id}$ loss functions as constraints, resulting in poorer accuracy in face alignment and reconstruction. This is because using a single attribute loss function leads to the loss of information from the other two attributes. The reason is that the network overly focuses on the supervised attribute during the learning process, thereby neglecting the variations in the other attributes. Consequently, the details and features of the other attributes are lost during the reconstruction process. Furthermore, different facial attributes are often correlated and exhibit mutual dependencies and associations. Using only single attribute loss functions, the interrelationships between facial attributes and their joint variations are not adequately modeled, overlooking the complexity of their relationships. Therefore, we introduced the $\mathcal{L}_{pose}$, $\mathcal{L}_{exp}$, and $\mathcal{L}_{id}$ loss functions in the fusion network, with the experimental results demonstrating a decrease in the normalized mean error in face alignment and reconstruction on the AFLW and AFLW2000-3D datasets. To further improve the alignment accuracy of the 3D face, we introduced the sparse 2D facial landmark loss $\mathcal{L}_{68}$ to enhance the geometric contour information during reconstruction. Remarkably, our full hybrid loss function achieved higher accuracy compared to other baselines on these two datasets.

3.2 A Visualization Experiment Performed on LS3D-W

This section compares the qualitative results of DRTN with the most appropriate state-of-the-art methods in the LS3D-W[5]. **Fig. 1** illustrates the 3D face reconstruction and landmark alignment results in complex scenes with insufficient illumination under a large facial pose. The red dashed box in the figure highlights that DRTN attains higher reconstruction accuracy than 3DDFA[3], DAMDNet[1], MARN[2], and MFIRRN[7] in high-fidelity 3D face reconstruction. This is more evident, especially regarding facial features contour, which validates that our model can better capture local details and geometric contour information. The LS3D-W [5] is collected in a high-level unconstrained complex environment under different illumination and face offset angles. The blue dashed box in the figure shows the face landmark alignment results, revealing that DRTN still demonstrates good alignment results under self-obscuring faces and low-light conditions. Therefore, the excellent experimental results verify the powerful 3D reconstruction capability and landmark alignment accuracy of the face attribute disentangling representation method.



Fig. 1 Comparison of models of different methods on the LS3D-W dataset. The red dashed box encloses the high-fidelity 3D face. The blue dashed box shows the 3D landmark alignment results.

3.3 Qualitative Results on CelebA and Casual Photos

To further validate the generalization ability of DRTN on other datasets, we used casual photos and the CelebA [4] for evaluation. The face images in the green dashed box in **Fig. 2** are the experimental results of CelebA [4]. The face images in the red dashed box are from some random photos of life and pictures

of anime characters. Our DRTN accomplishes accurate landmark alignment and fine 3D face reconstruction on both CelebA[4] and casual images, reflecting its good generalization ability. Since DRTN allows the network to learn the correlation between identity, expression, and pose attributes and fuse the underlying information between face attributes, accurate high-fidelity 3D faces, and 3D face landmark alignment can be reconstructed.



Fig. 2 Qualitative results on casual photos and CelebA. (a) original map (b) 3D landmark alignment results of the input. (c) high-fidelity 3D face (d) facial geometry.

References

- [1] 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).
- [2] 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.
- [3] 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).
- [4] 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).
- [5] 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 Proceedings of the IEEE International Conference on Computer Vision (pp. 1021-1030).
- [6] 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.
- [7] 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.