Supplementary of Disentangled Representation Learning for 3D Face Shape

1. Network Structure

Our network structure is shown in Fig. 1, and we choose Chebyshev polynomials of order 2 as hyper-parameter of our convolution layers. During training process, we duplicate identity and expression branch, respectively, to get the disentangling loss L_{dis} as given in Sec.3.4.

2. Latent space dimension exploration

In our paper, we compare our model ability with other baseline models on FaceWareHouse with latent space size is 25 for expression and 50 for identity. We also conduct experiment about our method with different size of latent space. The result shown in Tab. 1.

method	E_{avd}	E_{sed}	E_{id}	E_{exp}
Ours (25/10)	6.7/5.9	0.06/0.02	1.3/1.3	0.4/0.3
Ours (75/50)	3.7/2.8	0.02/0.00	0.9/0.9	0.3/0.2
Ours (50/25)	4.7/3.8	0.03/0.00	1.2/1.2	0.4/0.3

Table 1. More quantitative results. Ours$(25/10)$ represents that identity latent dim is set to 25 and expression latent dim is set to 10. So do Ours(75/50) and original result, Ours(50/25). All number in 0.1 millimeters.

3. Deformation Representation Reconstruction Accuracy

We use deformation representation in our framework, and the conversion from deformation representation to 3D mesh is solved by a least-square problem. We compute the geometric distance between original point clouds and DR-reconstructed ones over FaceWarehouse, and the average error is 31 micrometers. It means that the conversation process has very little influence on reconstruction accuracy.

4. Data Augmentation Samples

As descripted in Sec.3.4, we augment 2000 meshes with neutral expression from the FaceWareHouse dataset for identity decomposition branch training, and Fig. 2 shows some examples from the augmented models.

5. COMA Dataset [1]

5.1. Selected Expressions from COMA Dataset

In Fig 3, we show our selected 144 expressions from COMA dataset [1] for our decomposition and fusion networks pretraining in Sec.4.4. Each column is of the same identity with 12 various expressions.

5.2. 12 Cross Validation Experiments Result

We show the numerical result of 12 cross validation experiments compared with FLAME [2] in Tab 2. Our method gets lower error in most cases. For some case like bareteeth, our method gets higher median error than FLAME. Most error of our method is caused by the bias resulting from manual selection on expressions.

References

[1] S. S. Anurag Ranjan, Timo Bolkart and M. J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, 2018. 1

[2] T. Li, T. Bolkart, M. J. Black, H. Li, and J. Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 1, 5

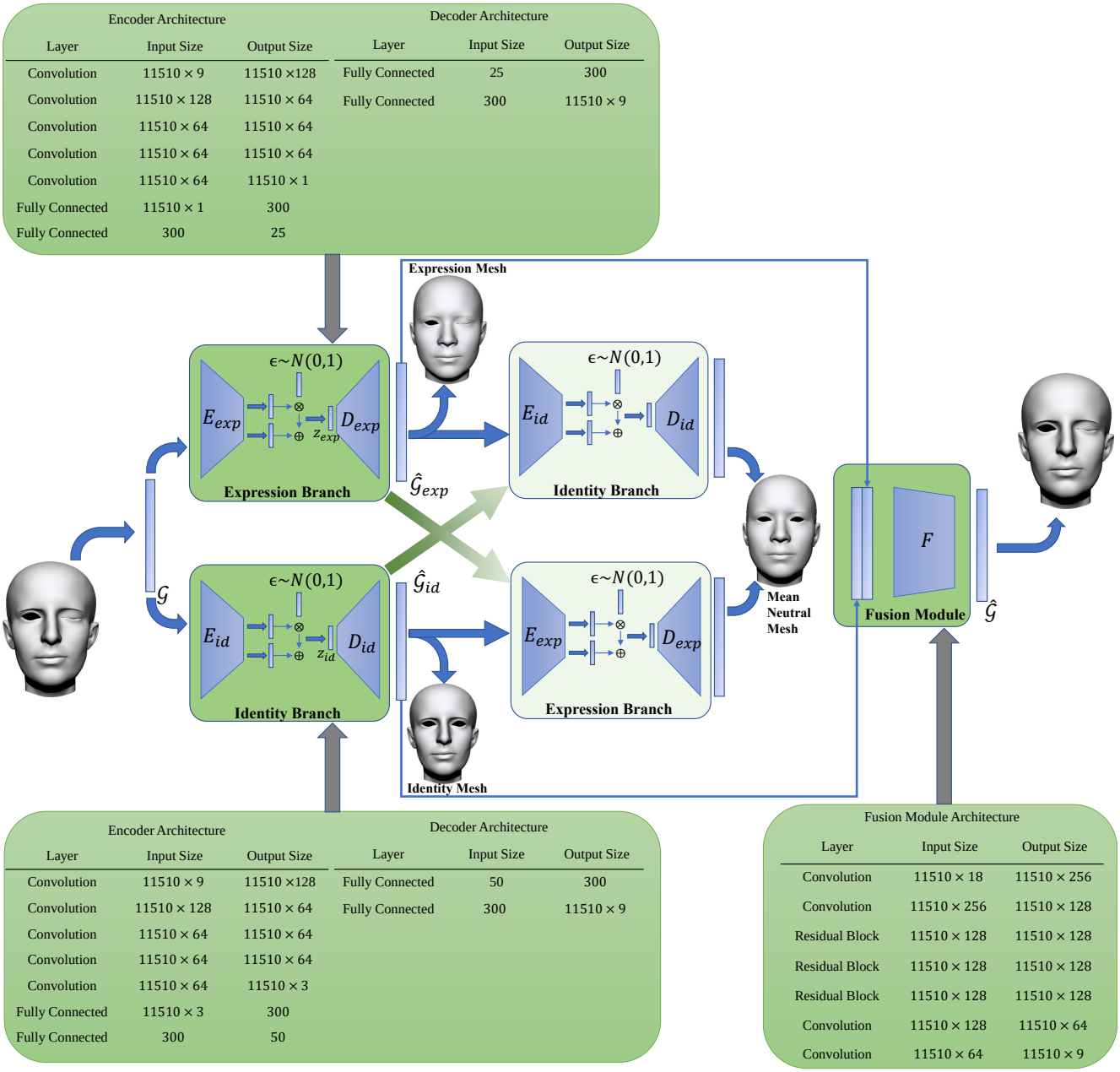

Figure 1. Our Network Structure.

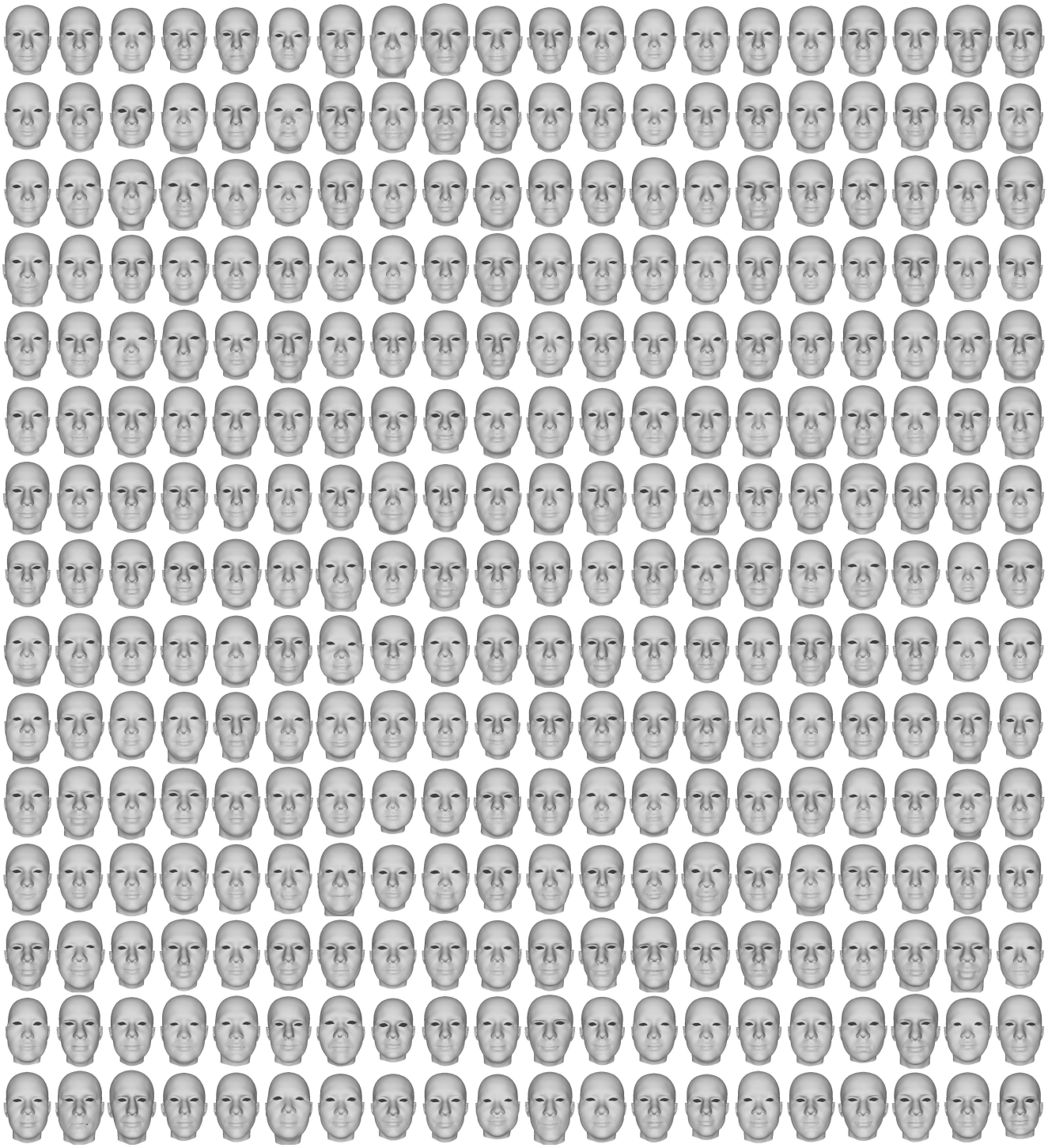

Figure 2. Data augmentation samples.

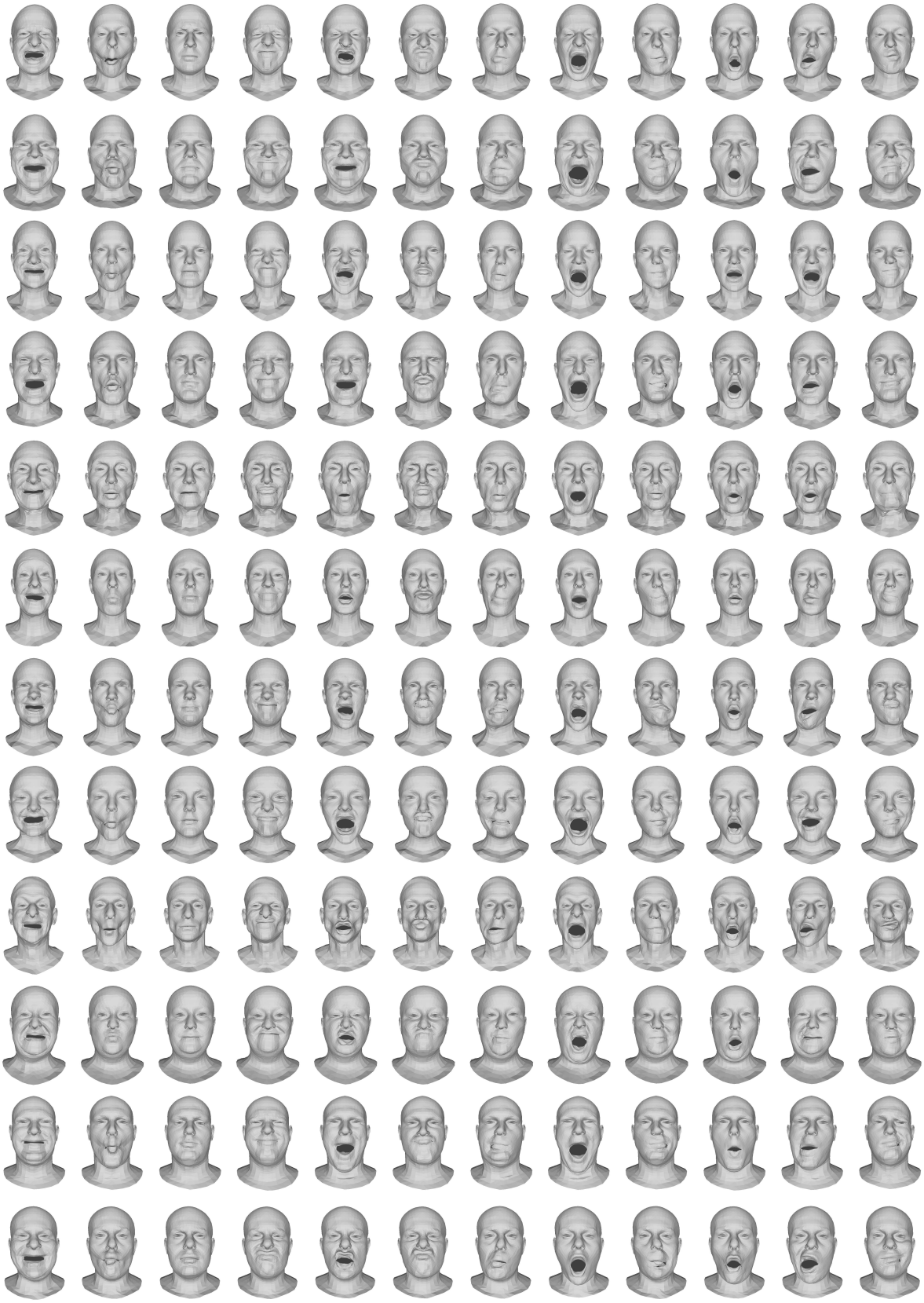

Figure 3. Selected 144 expressions from COMA dataset.

	Ours		FLAME [2]	
	Mean Error	Median	Mean Error	Median
bareteeth	**1.695**	1.673	2.002	**1.606**
cheeks in	**1.706**	**1.605**	2.011	1.609
eyebrow	**1.475**	**1.357**	1.862	1.516
high smile	**1.714**	1.641	1.960	**1.625**
lips back	**1.752**	**1.457**	2.047	1.639
lips up	**1.747**	**1.515**	1.983	1.616
mouth down	**1.655**	**1.587**	2.029	1.651
mouth extreme	**1.551**	**1.429**	2.028	1.613
mouth middle	**1.757**	1.691	2.043	**1.620**
mouth open	**1.393**	**1.371**	1.894	1.544
mouth side	**1.748**	**1.610**	2.090	1.659
mouth up	**1.528**	**1.499**	2.067	1.680
Average	**1.643**	**1.536**	2.001	1.615

Table 2. Comparison between our method and FLAME [2] on expression extrapolation experiment by testing on COMA dataset. Errors are in millimeters.