Generating Diverse 3D Reconstructions from a Single Occluded Face Image (*Supplementary Material*)

Rahul Dey Vishnu Naresh Boddeti
Michigan State University
East Lansing, MI
deyrahul,vishnu@msu.edu

The supplementary material is organized as follows:

1. We give further experimental results to support our claims regarding the effectiveness of the proposed method in Sec. 1.
2. We provide further quantitative and qualitative results, respectively, to evaluate the effectiveness of Diverse3DFace in Secs. 1.1 and 1.7.
3. We analyze the shape fitting performance on occluded images by FLAME [6] and our global+local models using a histogram of MSE errors in Sec. 1.2.
4. We study how the diversity hyperparameters affect the diversity and quality of the generated 3D reconstructions in Sec. 1.3
5. We present examples of diverse 3D reconstructions by our method on real-world occlusions in Sec. 1.4.
6. We study how occlusions at different regions of the face affect diverse 3D reconstruction differently in Sec. 1.5.
7. We study a controlled way of generating diverse reconstructions using interpolation in the latent space in Sec. 1.6.
8. Finally, we describe the full implementation details of our optimization routine in Sec. 2.1, and of the Mesh-VAE in Sec. 2.2.

1. Further Experiments

1.1. Further Quantitative Analysis on Diversity

We provide further quantitative evaluation of our approach compared to the baselines in terms of diversity performance as measured by the proposed *ASD-O*, *ASD-V* metrics, and the ratio *ASD-O/ASD-V*, on the CelebA dataset [7]. Since the CelebA dataset [7] is not labeled with groundtruth 3D shape, we do not compute the Closest Sample Distance (*CES*) on this dataset. To re-iterate, lower *ASD-V* indicates better consistency with the visible regions; and higher *ASD-O* indicates higher diversity in the occluded regions. As reported in Tab. 1, our approach obtains the maximum *ASD-O* across all occlusion types, the lowest *ASD-V* for *Glasses*, as well as the second lowest (compared to Mesh-VAE) *ASD-V* for *Facemasks* and *Random* occlusions. This is further corroborated by the significantly higher *ASD-O/ASD-V* ratios reported by Diverse3DFace compared to the baselines. Compared to this, single-stage diversity fitting baselines *viz.* FLAME+DPP and Global+Local+DPP generate the lowest *ASD-O/ASD-V* ratios, signifying that the 3D reconstructions generated by these approaches are neither diverse on the occluded regions, nor consistent with respect to the visible regions. On the other hand, one-pass samples generated by Global+Local+VAE are consistent with the visible face as reported by low *ASD-V*, but not diverse on the occluded regions (low *ASD-O*).

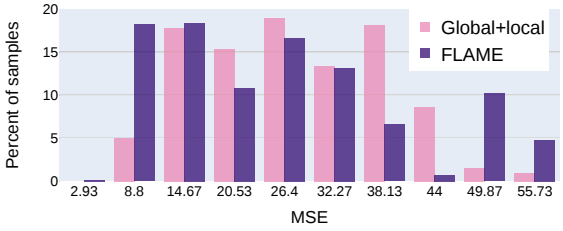

Figure 1. Histogram of MSE for shape fitting on occluded face images by FLAME [6] and our Global+local model.

1.2. Error Histogram Analysis

In Fig. 1, we plot the histograms of shape fitting errors (in terms of MSE) when the FLAME [6] and our global+local model are used to fit to partially occluded face images. One can observe that, while FLAME registers smaller errors (less than 10 MSE) on more number of

Occlusion Type	FLAME+DPP			Global+Local+DPP			Gloal+Local+VAE			Diverse3DFace (Ours)		
	ASD-V (↓)	**ASD-O** (↑)	$\frac{\text{ASD-O}}{\text{ASD-V}}$ (↑)	**ASD-V** (↓)	**ASD-O** (↑)	$\frac{\text{ASD-O}}{\text{ASD-V}}$ (↑)	**ASD-V** (↓)	**ASD-O** (↑)	$\frac{\text{ASD-O}}{\text{ASD-V}}$ (↑)	**ASD-V** (↓)	**ASD-O** (↑)	$\frac{ASD-O}{ASD-V}$ (↑)
Glasses	3.44	2.98	0.866	2.15	2.99	1.391	0.81	1.17	1.444	**0.68**	**3.56**	**5.235**
Face-mask	3.45	4.93	1.429	2.85	3.99	1.400	**0.75**	1.62	2.160	1.03	**7.47**	**7.252**
Random	4.12	4.23	1.027	3.17	3.84	1.211	**0.79**	1.29	1.633	0.83	**4.30**	**5.181**
Overall	3.86	4.44	1.150	3.03	3.88	1.281	**0.78**	1.41	1.808	0.90	**5.41**	**6.011**

Table 1. Quantitative evaluation of the diversity in 3D reconstruction of occluded faces from the CelebA dataset [7] between the baselines *vs.* Diverse3DFace in terms of ASD-V and ASD-O metrics (in order of 10^{-3}) and the ratio between them.

k \ n_σ	**1**	**2**	**3**	**4**	**5**
0.1	0.53	0.81	0.93	1.40	1.88
0.25	0.69	0.95	1.18	1.61	1.98
0.5	0.86	1.02	1.30	1.94	2.14
1	0.81	1.05	1.23	1.92	2.03
2	0.79	0.98	1.06	1.57	1.98

(a) ASD-V (↓)

k \ n_σ	**1**	**2**	**3**	**4**	**5**
0.1	3.63	4.92	5.62	7.17	8.64
0.25	4.13	6.37	7.65	8.18	10.73
0.5	5.98	8.25	9.16	11.19	14.53
1	5.18	7.89	8.84	10.72	12.96
2	4.42	6.68	7.40	9.78	12.21

(b) ASD-O (↑)

Table 2. Effect of the hyperparameters k and n_σ on the diversity metrics *ASD-V* and *ASD-O* on the CoMA dataset [9].

samples than the global+local model, there are significantly more number of samples ($\sim 15\%$) where FLAME registers very high MSE errors (> 50 MSE) than the global+local model. One can conclude that our global+local model is more robust than the global FLAME model [6] on samples with challenging occlusions.

1.3. Diversity Hyperparameters

The diversity generated by our approach is determined by the DPP loss $L_{dpp} = -tr\left(\mathbf{I} - (\mathbf{L} + \mathbf{I})^{-1}\right)$. Here,

Target Image | Fitting by Global-local model | 3D Reconstructions by Diverse3DFace

Figure 2. Set of 3D reconstructions by Diverse3DFace on real-world occluded face images.

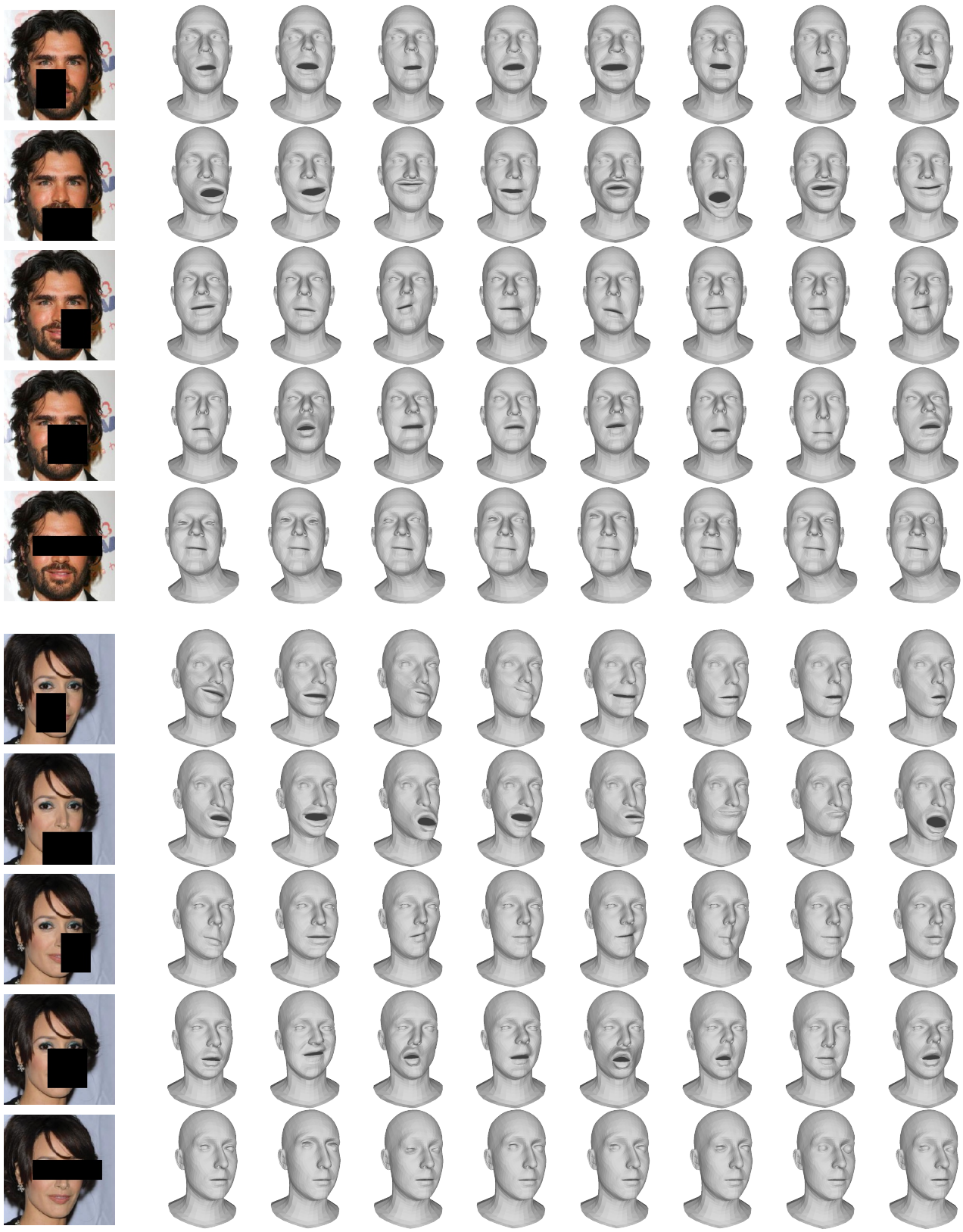

Figure 3. Qualitative evaluation of the diversity and robustness performance of Diverse3DFace to occlusions at different facial locations.

Figure 4. **Controlled generated of diverse 3D reconstructions between two distinct modes.** Diverse3DFace can be used to generate controlled diversity on the occluded regions by performing interpolation between two distinct shapes in the latent space.

the DPP kernel entry for the i,j-th element is given by $L_{i,j} = q_i S_{i,j} q_j$, where q_i denotes the quality of element i, and $S_{i,j}$ represents the similarity between i and j. The DPP optimization tries to maximize the quality of each sample, while minimizing the similarity between distinct samples. As stated in the main paper, we control the similarity term $S_{i,j} = \exp\left(-\frac{k}{\text{med}_{i,j}(dist_{i,j})} dist_{i,j}\right)$ and the quality term $q_i = \exp(-\max(0, \mathbf{z}_i^T \mathbf{z}_i - n_\sigma \sqrt{d}))$ using two parameters k and n_σ, respectively. In Tab. 2, we study the effects of the two hyper-parameters k and n_σ on diversity as measured by the diversity metrics *ASD-V* and *ASD-O*. As shown in Tab. 2, we obtain maximum *ASD-V*, as well as, *ASD-O* at $k = 0.5$; whereas both metrics increase as n_σ increases. Thus, we set $k = 0.5$ in our experiments while we choose $n_\sigma = 3$ as a sweet spot between minimizing *ASD-V* and maximizing *ASD*-O. The user can change the value of n_σ to tweak the diversity-realism trade-off.

1.4. Real-world Occlusions

We present examples of diverse 3D reconstructions by our approach on real-world occluded face images in Fig. 2. For these images, we inferred the occlusion mask using the face segmentation model by Nirkin *et al.* [8]. These results further demonstrate the efficacy of Diverse3DFace to generate diverse, yet plausible 3D reconstructions on real world occlusions ranging from glasses, scarf, facemasks, *etc*.

1.5. Moving the Occlusion Around the Face

In this section, we evaluate the diversity and robustness performance of Diverse3DFace to occlusions at different locations on the face. Fig. 3 shows the set of 3D reconstruction by Diverse3DFace when the occlusion moves around the face occupying the left cheek, mouth, the right cheek, center and the periocular (eye) regions of the face. Our method generates diverse, yet plausible set of 3D reconstructions for all the cases. We particularly note the high degree of diversity in expression that occurs when the mouth region is occluded, as is expected.

1.6. Diversity Interpolations

A potential application of Diverse3DFace is to perform controlled diversification around an occluded region during 3D reconstruction. To do this, we can first generate a set of diverse 3D reconstructions for an occluded target image and then allow the user to select two distinct samples to perform interpolation in-between. We perform interpolation in the latent space: $\mathbf{z}(\alpha) = \alpha \mathbf{z}_1 + (1-\alpha)\mathbf{z}_2$. This affords the user control over the extent and type of diversity. We present examples of such interpolations in Fig. 4.

1.7. Further Qualitative Results on CelebA Dataset

We show further qualitative results of diverse 3D reconstructions on occluded face images from the CelebA dataset [7] by Diverse3DFace, compared to the singular reconstruction by FLAME [6], DECA [4], CFR-GAN [5],

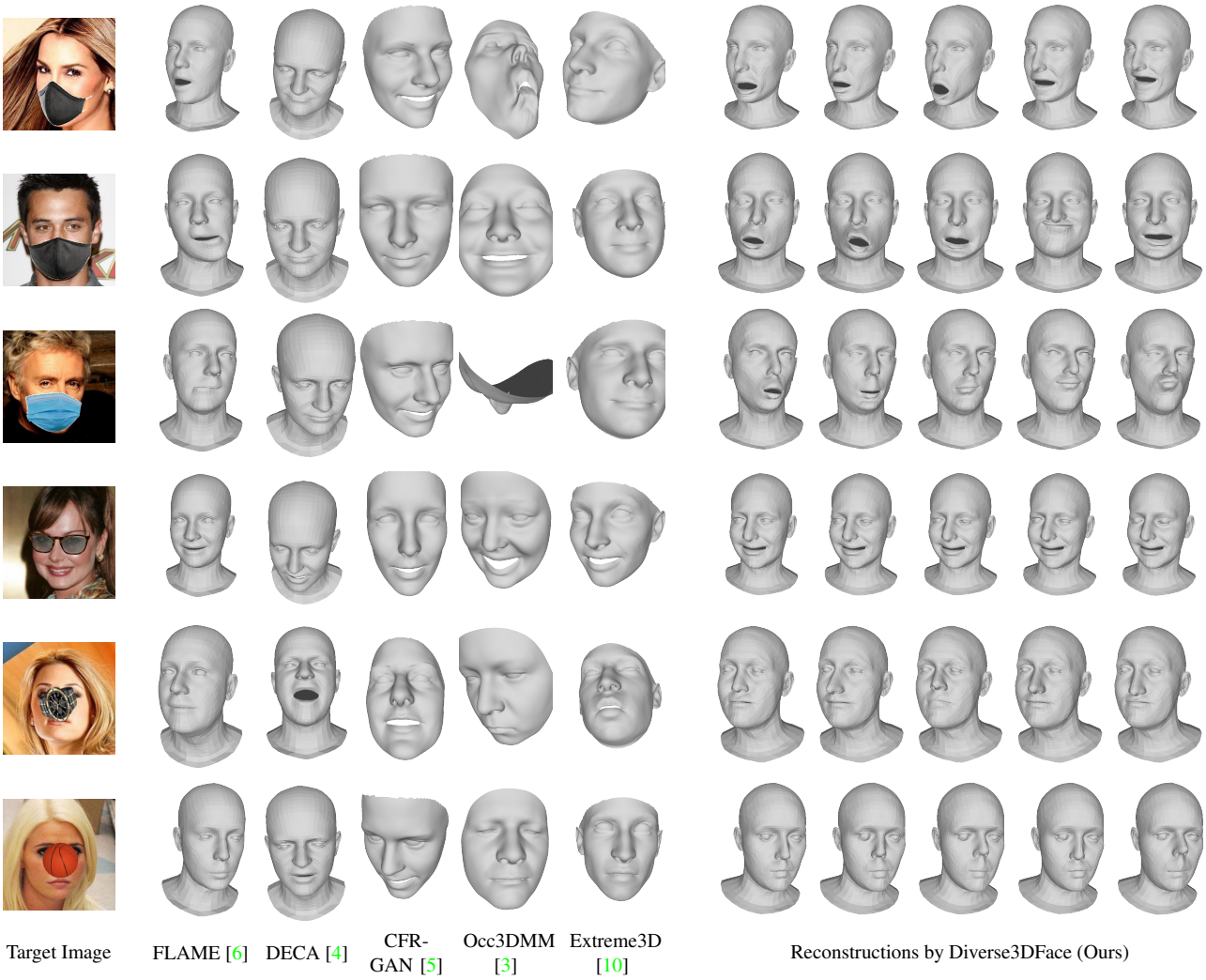

Figure 5. **More Qualitative evaluation on the CelebA dataset [7]**: Reconstructed singular 3D meshes from the target image by the baselines *vs.* the diverse reconstructions from Diverse3DFace.

Occ3DMM [3] and Extreme3D [10] in Fig. 5. While the baselines often get the pose, shape or expression wrong, Diverse3DFace generates 3D reconstructions that are consistent with the visible regions, yet plausibly diverse on the occluded regions.

2. Implementation Details

2.1. Optimization

We use the *PyTorch* library to implement our approach. In our experiments, we found that the SGD optimizer, with a learning rate of 5×10^{-3} gives the best results as compared to the Adam and RMSprop optimizers. For photometric fitting, we used the texture model provided by FLAME. We run the fitting stage (Algorithm 1) for $n_{iter} = 2000$ iterations and the diversity stage (Algorithm 2) for $n_{comp} = 300$ iterations. In Algorithm 1, we set the loss weights as follows: $\lambda_1^f = 5, \lambda_2^f = 16, \lambda_3^f = 10^{-3}$. During the diversifying shape completion stage (Algorithm 2), we set $\lambda_1 = 1000, \lambda_2 = 500, \lambda_3 = 0.025$. Further, we found that using a slightly smaller learning rate for the eyeball components while fitting the global+local model gives better results. For these components, we set the learning rate to be 0.5 times that of the other components.

2.2. Mesh-VAE

The Mesh-VAE model is based on the fully convolutional mesh autoencoder (Meshconv) architecture proposed by Zhou *et al.* [11]. Meshconv [11] uses spatially varying convolutional kernels for different mesh vertices to account

Input	Layer	Output size	Output
5023×3 **Mesh**	$\rightarrow$ vcDownConv($in_c = 3, out_c = 32, s = 2, r = 43, M = 17$) + vcDownRes(2)	1367×32	
	vcDownConv($in_c = 32, out_c = 64, s = 1, r = 27, M = 17$) + vcDownRes(1)	1367×64	
	vcDownConv($in_c = 64, out_c = 128, s = 2, r = 54, M = 17$) + vcDownRes(2)	270×128	
	vcDownConv($in_c = 128, out_c = 256, s = 1, r = 25, M = 17$) + vcDownRes(1)	270×256	
	vcDownConv($in_c = 256, out_c = 512, s = 2, r = 81, M = 17$) + vcDownRes(2)	45×512	
	vcDownConv($in_c = 512, out_c = 1024, s = 1, r = 27, M = 17$) + vcDownRes(1)	45×1024	*feats*
feats	vcDownConv($in_c = 1024, out_c = 64, s = 2, r = 37, M = 17$) + vcDownRes(2)	10×64	$\boldsymbol{\mu}$
feats	vcDownConv($in_c = 1024, out_c = 64, s = 2, r = 37, M = 17$) + vcDownRes(2)	10×64	$\log \boldsymbol{\sigma}^2$
Model Complexity	9M		

Table 3. Network architecture of the Mesh-VAE Encoder $\mathcal{E}_{mesh}$.

Input	Layer	Output size	Output
10×64 **z**	vcUpConv($in_c = 64, out_c = 1024, s = 2, r = 8, M = 17$) + vcUpRes(2)	45×1024	
	vcUpConv($in_c = 1024, out_c = 512, s = 1, r = 27, M = 17$) + vcUpRes(1)	45×512	
	vcUpConv($in_c = 512, out_c = 256, s = 2, r = 16, M = 17$) + vcUpRes(2)	270×256	
	vcUpConv($in_c = 256, out_c = 128, s = 1, r = 25, M = 17$) + vcUpRes(1)	270×128	
	vcUpConv($in_c = 128, out_c = 64, s = 2, r = 12, M = 17$) + vcUpRes(2)	1367×64	
	vcUpConv($in_c = 64, out_c = 32, s = 1, r = 27, M = 17$) + vcUpRes(1)	1367×32	
	vcUpConv($in_c = 32, out_c = 3, s = 2, r = 24, M = 17$) + vcUpRes(2)	5023×3	**Output**
Model Complexity	8M		

Table 4. Network architecture of the Mesh-VAE Decoder $\mathcal{D}_{mesh}$.

for the irregular structure of a 3D mesh. The spatially varying kernels are sampled from the span of a shared weight basis, using learned per-vertex coefficients. In addition, Meshconv defines pooling and unpooling operations on a 3D mesh by performing feature aggregation Monte Carlo sampling [11].

We trained the Mesh-VAE with FLAME [6] registered groundtruth scans provided in the CoMA [9] and D3DFACS [2] datasets. We perturbed the input meshes with uniformly sampled rectangular masks (in XY) within a range around the mesh center, while gradually increasing the size of the mask per training epoch until it covered $\sim$40% of the vertices. We detail the network architecture for the Mesh-VAE in Tabs. 3 and 4.

The abbreviated operators used are defined as follows:

- vcDownConv(in_c, out_c, s, r, M) + vcDownRes(s): Downward residual block (as defined in Meshconv [11]), with in_c input channels, out_c output channels, s stride, r kernel radius and M number of shared weight bases. The output is activated with ELU [1] activation.

- vcUpConv(in_c, out_c, s, r, M) + vcUpRes(s): Upward residual block (as defined in Meshconv [11]), with in_c input channels, out_c output channels, s stride, r kernel radius and M number of shared weight bases. The output is activated with ELU [1] activation.

References

[1] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. 6

[2] Darren Cosker, Eva Krumhuber, and Adrian Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modeling. In *2011 international conference on computer vision*, pages 2296–2303. IEEE, 2011. 6

[3] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126(12):1269–1287, 2018. 5

[4] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 4, 5

[5] Yeong-Joon Ju, Gun-Hee Lee, Jung-Ho Hong, and Seong-Whan Lee. Complete face recovery gan: Unsupervised joint face rotation and de-occlusion from a single-view image. In *WACV*, 2022. 4, 5

[6] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 1, 2, 4, 5, 6

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 4, 5

[8] Yuval Nirkin, Iacopo Masi, Anh Tran Tuan, Tal Hassner, and Gerard Medioni. On face segmentation, face swapping, and face perception. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 98–105. IEEE, 2018. 4

[9] Anurag Ranjan, Timo Bolkart, Soubhik Sanyal, and Michael J. Black. Generating 3D faces using convolutional mesh autoencoders. In *European Conference on Computer Vision (ECCV)*, pages 725–741, 2018. 2, 6

[10] Anh Tuán Trán, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018. 5

[11] Yi Zhou, Chenglei Wu, Zimo Li, Chen Cao, Yuting Ye, Jason Saragih, Hao Li, and Yaser Sheikh. Fully convolutional mesh autoencoder using efficient spatially varying kernels. *arXiv preprint arXiv:2006.04325*, 2020. 5, 6