Supplemental: Learning Complete 3D Morphable Face Models from Images and Videos

Mallikarjun B R Ayush Tewari Hans-Peter Seidel Mohamed Elgharib Christian Theobalt
Max Planck Institute for Informatics, Saarland Informatics Campus

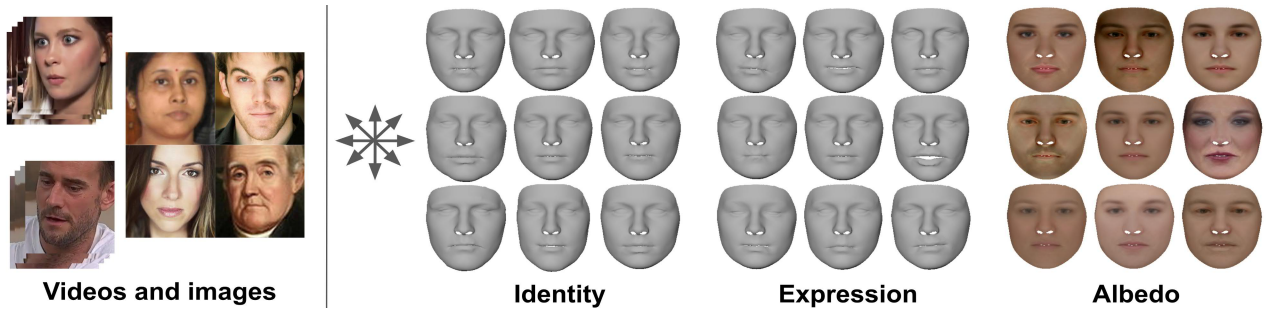

Figure 1. We present a method for learning complete 3D morphable models of faces from videos and images. We show visualizations of the learned models on the right. Faces in each direction of indicated arrows is obtained by linearly scaling individual component of respective models. Identity geometry captures variations in the face shape (second column), lips (top left to bottom right) and jaw (top right to bottom left), while expressions capture variations due to mouth opening (second row), smile (second column) and eye movement (top right to bottom left). Albedo/Reflectance spans a variety of skin color (second column), eye color (top right to bottom left) and gender specific features such as facial hair and make-up (second row).

In this supplemental document, we provide more training details, as well as more qualitative results, comparisons and ablative studies.

1. Vertex correspondences for landmark and segmentation consistency losses

We provide more details about computing the sliding contours on the mesh, used in the landmark and segmentation consistency loss terms. Unlike interior facial landmarks, we cannot annotate mesh vertices for the rolling contours (face contour and the inner contours of the lips). We compute these correspondences using our renderer. We first annotate several masks on the template mesh, see Fig 2. We use the boundaries between the projections of different regions to compute the rolling contour vertices. For eg.,the red and blue masks are used to compute the inner contour of the upper lip. Inner contour of the lower lip can be computed similarly. Face contour is computed as the boundary between the yellow mask and the background.

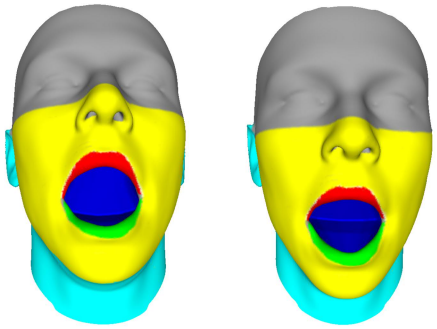

Figure 2. Masks used to compute the rolling contours.

2. Experimental details

Our network architecture is similar to that of FML [7]. A few convolutional layers extract features for each frame in the multi-frame image. These features are average pooled, and further processed using convolutional and fully connected layers to obtain the identity parameters. This ensures that the identity component of the reconstruction is consistent for the multi-frame input. Frame dependent parameters such as expression, illumination and rigid head pose are predicted independently for each frame using further siamese

	Layers	Activation Shape	Siamese	Output
Image (240,240,3)	Conv2D (kernel 11x11, stride 4) + ReLU	(60, 60, 96)	Yes	*unnamed*
↑	MaxPool (kernel 3x3, stride 2)	(29, 29, 96)	n/a	*unnamed*
↑	Conv2D (kernel 5x5, stride 1) + ReLU	(29, 29, 256)	Yes	*unnamed*
↑	MaxPool (kernel 3x3, stride 2)	(14, 14, 96)	n/a	*unnamed*
↑	Conv2D (kernel 3x3, stride 1) + ReLU	(14, 14, 384)	Yes	lowFeatures
↑	Conv2D (kernel 3x3, stride 2) + ReLU	(7, 7, 256)	Yes	*unnamed*
↑	Conv2D (kernel 3x3, stride 2) + ReLU	(4, 4, 256)	Yes	mediumFeatures

Table 1. Feature extractor details.

	Layers	Activation Shape	Siamese	Output
mediumFeatures	Concat	(M, 4, 4, 256)	n/a	*unnamed*
↑	MeanPool	(4, 4, 256)	n/a	*unnamed*
↑	Conv2D (kernel 3x3, stride 1) + ReLU	(4, 4, 384)	No	*unnamed*
↑	Conv2D (kernel 3x3, stride 1) + ReLU	(4, 4, 256)	No	*unnamed*
↑	Fully Connected + ReLU	(1000, 1)	No	*unnamed*
↑	Fully Connected + ReLU	(1000, 1)	No	*unnamed*
↑	Fully Connected	(80+80, 1)	No	shapeParam + reflectanceParam

Table 2. Shared identity details.

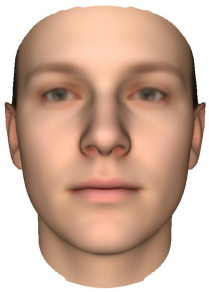

Figure 3. Mean mesh with albedo.

convolutions and fully connected layers on the per-frame features. Our expression model is represented using the deformation graph, while FML [7] uses a pre-trained model at the mesh resolution. For more details on the architecture of our network, please refer to Tables. 1, 2 and 3. Fig 3 shows the fixed mean mesh with albedo used by our method.

Empirically, we found the following weights help in stable training of our method:
$\lambda_{\text{pho}} = 2.5, \lambda_{\text{land}} = 50, \lambda_{\text{smo}} = 10, \lambda_{\text{seg}} = 0.001, \lambda_{\text{per}} = 1, \lambda_{\text{dis}} = 1$.

3. Results

In the following, we show more comparisons to state-of-the-art methods, and ablative studies on various loss terms.

3.1. Video results

In the project page[1], we provide visualization of our learned model. It can be observed that the learned model has good disentanglement of identity and expression components. In addition, we show expression transfer of videos to static faces. This also confirms the disentanglement quality of our model. Learning a personalized model helps in obtaining high quality reconstructions. Please watch the video in project page[1] to see comparitive results with FML. As FML uses a pre-trained expression model, it fails to capture various identity specific mouth movements, which are outside the expression model space. We show expression transfer results with personalized model to show that the semantics of the expressions are intact, even with personalized model.

3.2. Comparisons to state-of-the-art methods

Fig 4 shows personalized model reconstruction comparison with FML [7]. Note that the training strategy of FML does not allow for learning of the expression model. Thus, we obtain higher quality reconstructions. Please watch video.mp4 for video comparisons. Fig 6 shows several comparisons of our method with FML [7] and Tewari *et al*. [8]. Tewari *et al*. [8] learn a corrective space on top of the existing face models. The corrective space contains both identity and expression components, without disentanglement. FML [7] uses a pretrained expression model and hence can not generalize to various expressions. Figs 7 and 8 show more comparisons of our method with MoFA [9], GANFIT [1] and RingNet [5]. All these methods use pretrained geometry models. Fig 9 shows more comparisons of our method with Tran *et al*. [10], [11]. Richardson *et al*. [4] and Sela *et al*. [6] are trained on synthetic data and do not generalize well to real data (Fig. 10). These approaches cannot disentangle the identity and expression components.

[1] **project page:** http://gvv.mpi-inf.mpg.de/projects/LeMoMo/

	Layers	Activation Shape	Siamese	Output
shapeParam, reflectanceParam	Fully Connected + ReLU + Reshape	(14, 14, 1)	No	*unnamed*
↑	Conv2D (kernel 3x3, stride 1) + ReLU	(14, 14, 384)	No	*unnamed*
↑, lowFeatures	Concat	(14, 14, 768)	n/a	*unnamed*
↑	Conv2D (kernel 3x3, stride 1) + ReLU	(14, 14, 384)	Yes	*unnamed*
↑	Conv2D (kernel 3x3, stride 1) + ReLU	(14, 14, 384)	Yes	*unnamed*
↑	Conv2D (kernel 3x3, stride 1) + ReLU	(14, 14, 256)	Yes	*unnamed*
↑	MaxPool(kernel 3x3, stride 2)	(6, 6, 256)	Yes	*unnamed*
↑	Fully Connected + ReLU	(2048, 1)	Yes	*unnamed*
↑	Fully Connected	(6+64+27+1, 1)	Yes	pose + expressionParam + illuminationParam

Table 3. Parameter estimation details.

We are the only approach among all these approaches which learns the complete 3D morphable model without using any 3D supervision.

3.3. Ablative studies

Perceptual and segmentation loss Fig 5 shows reconstructions for several images with and without the perceptual loss. It is clearly evident that the perceptual loss helps in capturing detailed albedo and more realistic overlays. Fig 11 shows reconstructions for several images with and without the segmentation consistency loss. This loss helps in better capturing mouth shape and inner lip contours.

Orthogonality Orthogonality between the identity and expression models is ensured by design, as in FML. The network updates the identity model in each training iteration by projecting it onto the orthogonal complement of the expression model. Although theoretically, it should help in better disentanglement between the identity and expression components, empirically we found that it does not lead to any significant difference. We report the disentanglement (average of expression deformation for images with neutral faces) and reconstruction errors proposed in the main paper in Tab. 5 and Tab. 4. While the disentanglement loss becomes slightly higher, reconstruction error decreases with orthogonality.

Disentanglement In this section, we show the impact of identity pre-training explained the main paper. As shown in Tab. 5, pre-training improves reconstructions in terms of the disentanglement metric.

3.4. Fitting to 3D scans

We also evaluate our learned models by fitting them to 3D scans of the BU-3DFE dataset [12]. Here we use PCA model with 100 basis vectors each for both identity and expression, learnt from 10000 images from VoxCeleb dataset. We use the dense correspondences precomputed for quantitative evaluations, and minimize the per-vertex distance between the reconstructed and ground-truth meshes. The translation and scale of the reconstructed mesh is computed at the first iteration and fixed during optimization. ADAM [2] optimizer with a step size of 0.05 is used to minimize the per-vertex distances, as well as the global rotation of the reconstructed mesh. Table 6 reports the errors averaged across 166 meshes of BU-3DFE dataset. Our method obtains low fitting errors. FML [7] obtains slightly lower errors in this evaluation, which does not look at the quality of image reconstructions.

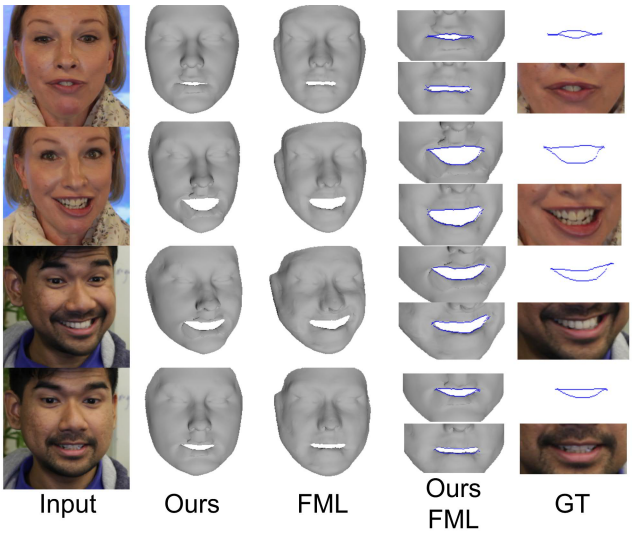

Figure 4. Our personalized model captures higher quality mouth geometry compared to FML, where only the identity models can be personalized. We show the inner contours of the meshes (ours-top, FML-bottom) in column 4. The ground truth inner contours and zoomed in image are visualized in column 5.

In Table 7, we evaluate the compactness of the learned models by reporting errors achieved by smaller-sized models. Our model achieves good quality even with 5 basis vectors.

3.5. Improvement over segmentation networks

Fig 12 shows lip regions as projected from the reconstructed mesh, compared to the lip segments predicted by the network [3]. Our method is robust to images which are of low quality and with faces in extreme poses.

	w/ O	w/o O	MoFA	FML	Fine [8]	Coarse [8]
Mean	**1.75**	1.78	3.22	1.78	1.83	1.81
SD	0.44	**0.43**	0.77	0.45	0.39	0.47

Table 4. Geometric reconstruction error (in mm) on the BU-3DFE dataset [12]. Our technique outperforms *MoFA* [9], coarse and fine models of Tewari *et al*. [8] and *FML et al*. [7].

	w/o $\mathcal{L}_{dis}$	w/o PT	w/o O	Ours	FML	MoFA
AE: Mean	2.5075	0.0147	**0.0105**	0.0116	2.0329	0.4056
AE: SD	0.8290	0.0464	0.0506	**0.0385**	0.6840	0.2

Table 5. Our identity disentanglement term results in lesser leakage of identity geometry into expression component in neutral faces. We observe slightly worse disentanglement without identity pre training (PT). Our method performs better than *FML* [7] and *MoFA* [9]

Figure 5. Albedo and overlay is noticeably improved with the perceptual loss

	FML	Ours
3D error	0.76(σ=0.11)	1.07(σ=0.17)

Table 6. Geometric reconstruction error (in mm) on the BU-3DFE dataset [12], when the learned models are fit to 3D scans. The PCA models are used for fitting here.

4. Limitations

Although our method is robust to various aspects of in-the-wild images, it is still limited in certain cases. Fig 13 shows various limitations of our method. Our method fails to capture mid- and high-frequency details in the geometry component, such as wrinkles and beard. This is common with other model-based reconstruction approaches. It also fails to capture specularities and extreme lighting conditions as our model assumes a lambertian surface and distant illumination. In cases of occlusions, such as glasses, our method incorrectly compensates by reconstructing these details in the albedo component.

References

[1] Baris Gecer, Stylianos Ploumpis, Irene Kotsia, and Stefanos Zafeiriou. Ganfit: Generative adversarial network fitting for high fidelity 3d face reconstruction. In *CVPR*, 2019. 2, 6

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015. 3

[3] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. *arXiv preprint arXiv:1907.11922*, 2019. 3, 8

[4] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *CVPR*, July 2017. 2

[5] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and

Figure 6. Our approach produces better geometry, including detailed mouth shapes compared to Tewari *et al*. [9] and FML [7]. Our albedo is also more detailed and better disentangled from the illumination.

expression from an image without 3D supervision. In *CVPR*, pages 7763–7772, 2019. 2, 7

Figure 7. MoFA [9] and GANFIT [1] produce less accurate mouth shapes than our technique. GANFIT [1] can produce artifacts in the albedo and final overlay, especially around the eyes.

[6] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. In *International Conference on Computer Vision (ICCV)*, 2017. 2

[7] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhoefer, and Christian Theobalt. FML: Face model learning from videos. In *CVPR*, 2019. 1, 2, 3, 4, 5

[8] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Flo-

	w/ 5	w/ 10	w/ 20	w/ 40	w/ all
3D error	1.71(σ=0.44)	1.62(σ=0.42)	1.34(σ=0.30)	1.11(σ=0.22)	1.07(σ=0.17)

Table 7. Compactness evaluation. Our learned 3DMM obtains low reconstruction errors, even with a subset of the bases. Geometric reconstruction error (in mm), with standard deviation (in brackets) on the BU-3DFE dataset [12], when the learned models are fit to 3D scans. The PCA models are used for fitting here. The numbers in the first row indicates number of basis vectors used for identity and expression models.

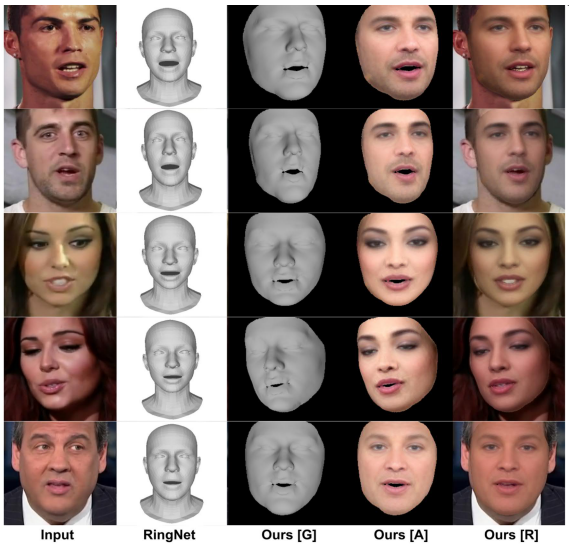

Figure 8. Our technique better captures mouth shape and eye geometry than *RingNet* [5]. It also produces a photorelistic overlay.

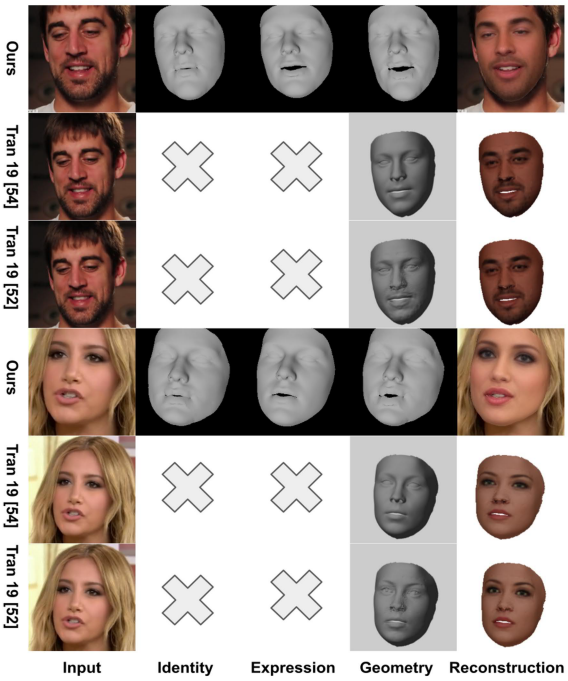

Figure 9. Both approaches of Tran *et al*. [11, 10] do not disentangle identity geometry from expressions. Our technique, however, estimates and disentagles all facial components. It also produces produces more accurate mouth shapes.

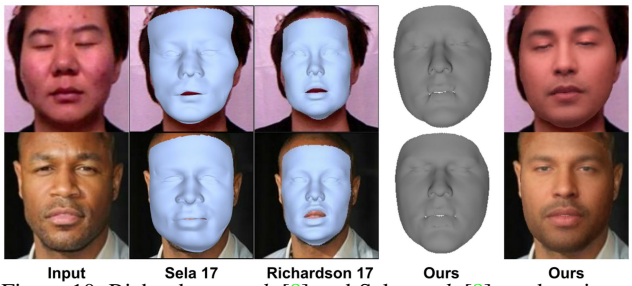

Figure 10. Richardson *et al*. [8] and Sela *et al*. [8] produce inaccurate geometry and do not estimate albedo nor illumination. Our approach estimates all facial components, including high quality geometry and overlay.

rian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *CVPR*, 2018. 2, 4, 7

[9] Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, pages 3735–3744, 2017. 2, 4, 5, 6

[10] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *CVPR*, 2019. 2, 7

[11] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. June 2019. 2, 7

[12] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and M. J. Rosato. A 3d facial expression database for facial behavior research. In *International Conference on Automatic Face and Gesture Recognition (FGR06)*, pages 211–216, 2006. 3, 4, 7

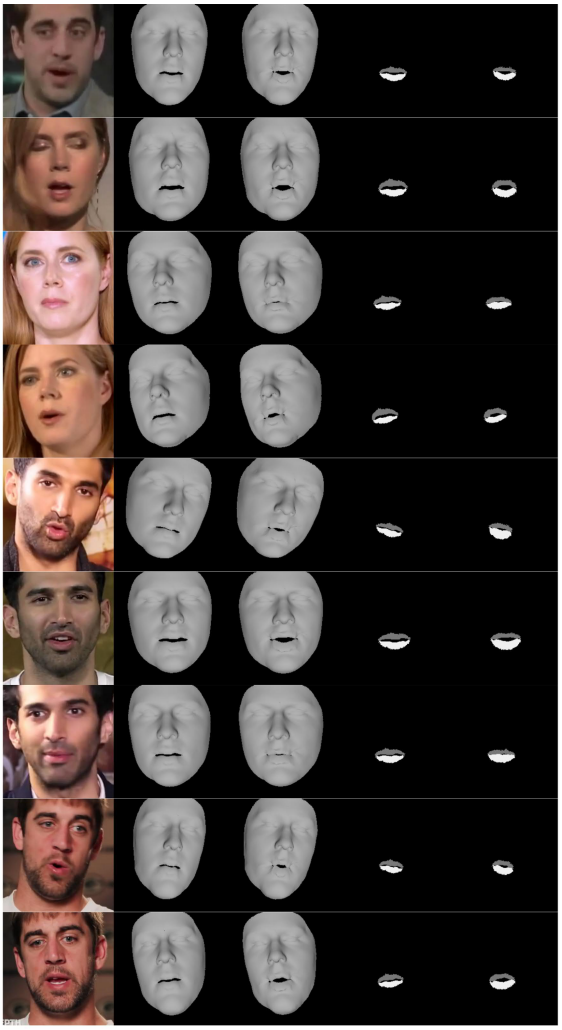

Figure 11. Observe that having segmentation consistency helps in better capturing mouth shape and expression.

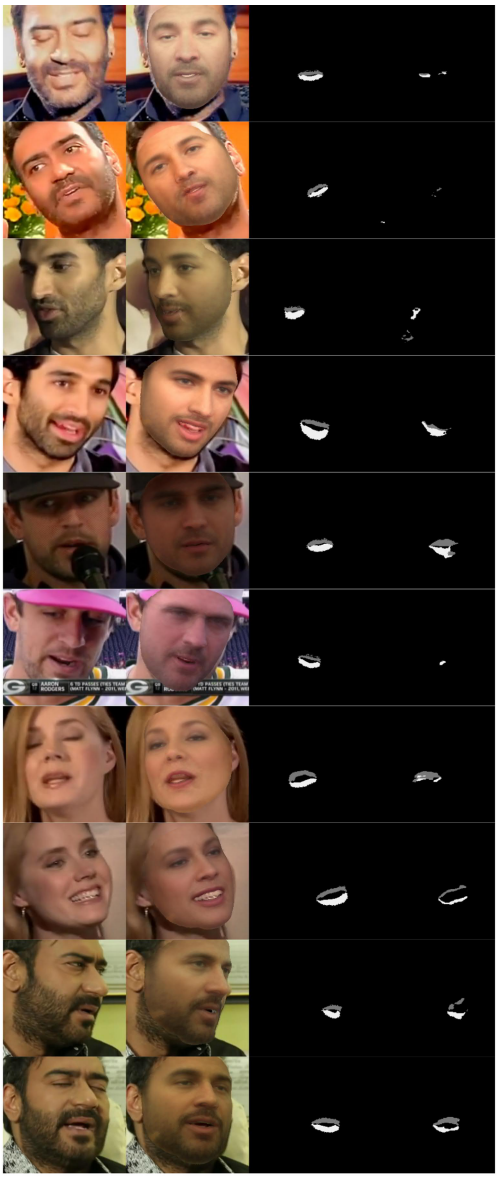

Figure 12. Our approach produces plausible upper (gray) and lower (white) lip segmentations even when the images are of bad quality, contain extreme poses or occlusions. In such cases [3] struggles to produce acceptable segmentation (see column 4).

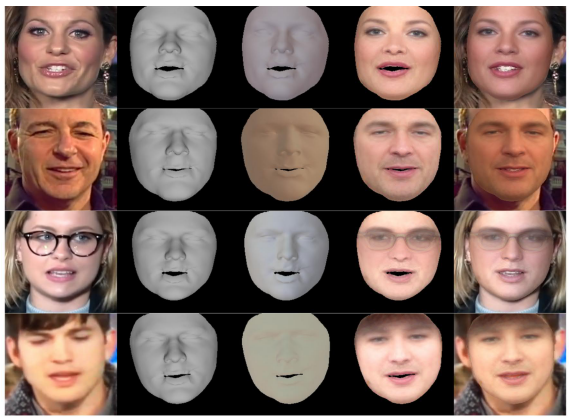

Figure 13. Limitations of our method. We cannot reconstruct high-frequency geometry details. In addition, we do not model specularities or occlusions.