# FML: Face Model Learning from Videos
# — Supplemental Document —

Ayush Tewari[1] Florian Bernard[1] Pablo Garrido[2] Gaurav Bharaj[2] Mohamed Elgharib[1]
Hans-Peter Seidel[1] Patrick Pérez[3] Michael Zollhöfer[4] Christian Theobalt[1]

[1]MPI Informatics, Saarland Informatics Campus [2]Technicolor [3]Valeo.ai [4]Stanford University


Figure 1. We propose multi-frame self-supervised training of a deep network based on in-the-wild video data for jointly learning a face model and 3D face reconstruction. Our approach successfully disentangles facial shape, appearance, expression, and scene illumination.

In this supplemental document, we provide more details on the network architecture and the empirically determined weights in the employed loss function. We also show more qualitative and quantitative comparisons and discuss limitations of our approach. In addition, we describe how to extract statistics from the learned model and visualize its modes. A large amount of qualitative results can also be found on the supplemental webpage[1].

## 1. Network Details

We provide further details of the feature extraction, shared identity and parameter estimation network in Tab. 1, 2 and 3, respectively. Overall, our network has 124M parameters. Note that a subset of these parameters includes the learned geometry and appearance model. We train our networks on commodity Titan Volta GPUs.

### 1.1. Visualizing the Modes of Variation

While the learned model is an optimal basis for the monocular face reconstruction task, it does not allow for an intuitive analysis of the most prominent modes of variation observed in the data. However, we can easily reparameterize the learned model and construct a new representation using e.g. Principle Component Analysis (PCA). More specifically, we compute PCA on 3D reconstructions obtained by our approach for over 10k images of our training set. Note, our approach is trained in a self-supervised manner without requiring ground truth in the form of dense geometry and appearance annotations. The new parametrization allows us to build a statistical face model of facial identity and appearance, as in [1], but based on in-the-wild video data, see Fig. 2, 3 and 4. Our model learns global variation modes that roughly correspond to gender (see Fig. 2) as well as local variation modes, such as nose and eye deformations (see Fig. 3). The visualization also shows the separation between the learned shape identity model and the expression model.

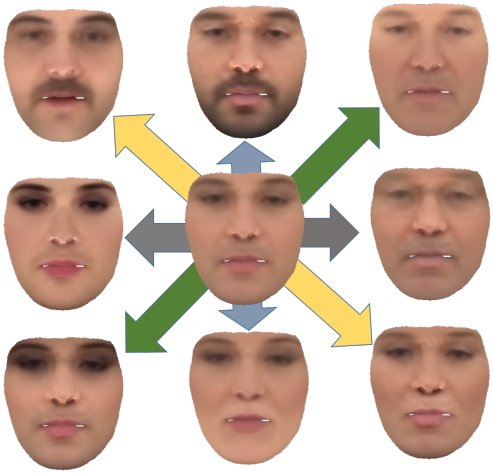

Figure 2. Visualization of the Reflectance Model. We show four of the learned modes.

## 2. Weights of the Energy

We found the weights in our energy empirically and kept them fixed in all experiments: $\lambda_{\text{pho}} = 1.6/|\tilde{\mathcal{V}}|, \lambda_{\text{lan}} =$

[1] http://gvv.mpi-inf.mpg.de/projects/FML19

Table 1. Feature Extractor Network Details. ↑ means that the input is taken from the layer in the row above.

| Input | 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 | (14, 14, 256) | Yes | *unnamed* |
| ↑ | Conv2D (kernel 3x3, stride 1) + ReLU | (14, 14, 384) | Yes | $\text{lowFeatures}_f$ |
| ↑ | Conv2D (kernel 3x3, stride 2) + ReLU | (7, 7, 256) | Yes | *unnamed* |
| ↑ | Conv2D (kernel 3x3, stride 2) + ReLU | (4, 4, 256) | Yes | $\text{mediumFeatures}_f$ |

Table 2. Shared Identity Network Details. ↑ means that the input is taken from the layer in the row above.

| Input | Layers | Activation Shape | Siamese | Output |
|---|---|---|---|---|
| $\text{lowFeatures}_{0\ldots M}$ | 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 | (500 + 500, 1) | No | shapeParam + reflectanceParam |

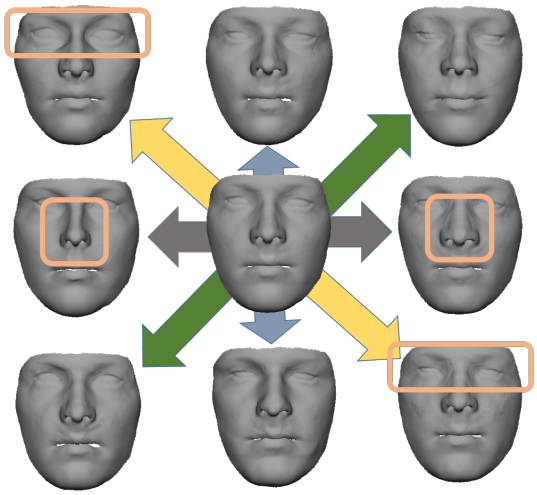

Figure 3. Visualization of the Shape Model. We show four of the learned modes.

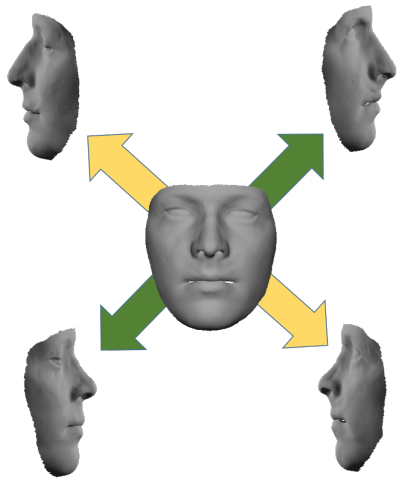

Figure 4. Visualization of the Shape Model. We show two of the learned modes from a side view.

$4.7, \lambda_{\text{smo}} = 0.001, \lambda_{\text{spa}} = 1e{-}7, \lambda_{\text{ble}} = 1e{-}8.$

## 3. Results

In the following, we show more detailed qualitative and quantitative evaluations of our approach.

**Monocular vs multi-frame reconstruction** Figure 5 shows the advantage of using multiple frames at test time. Multi-frame reconstruction improves overall consistency of the estimated 3D face and resolves ambiguity due to occlusions that are present in one of the images. Such ambiguities cannot be resolved completely when feeding only a single image to the network. Still, our network is able to obtain plausible facial identity for the monocular case thanks to our multi-frame based training.

**Comparison to state-of-the-art methods** Fig. 6, 7, 8 and 9 show more comparisons to related state-of-the-art approaches for monocular 3D face reconstruction. Our multi-frame based training succeeds in reconstructing 3D faces for images with large poses and harsh yet low-frequency illumination, as shown in Fig. 6. The method of Tewari et al. [8] is unable to deal with these cases, as it is trained on monocular images. Fig. 7 shows that our approach also generalizes well to facial identities having beards and non-average faces thanks to the learning of the optimal model from in-the-wild data. On the contrary, methods relying on synthetic data [5, 7] and/or an underlying 3DMM [9], fail to generalize to novel identities not explained by the 3D model or training corpus. Our approach not only estimates 3D faces from challenging in-the-wild images, but

Table 3. Parameter Estimation Network Details. ↑ means that the input is taken from the layer in the row above.

| Inputs | 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$_f$ | 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) | Yes | rigid$_f$ + expressionParam$_f$ + illuminationParam$_f$ |

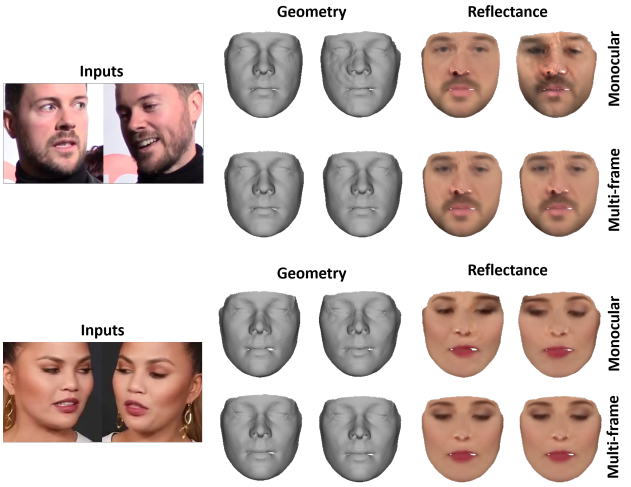


Figure 5. Monocular vs. multi-frame reconstruction. Multi-view reconstruction improves consistency and reconstruction quality, especially for regions occluded in one of the images. Note that all results are shown with a frontal pose and neutral expression for comparison purposes.

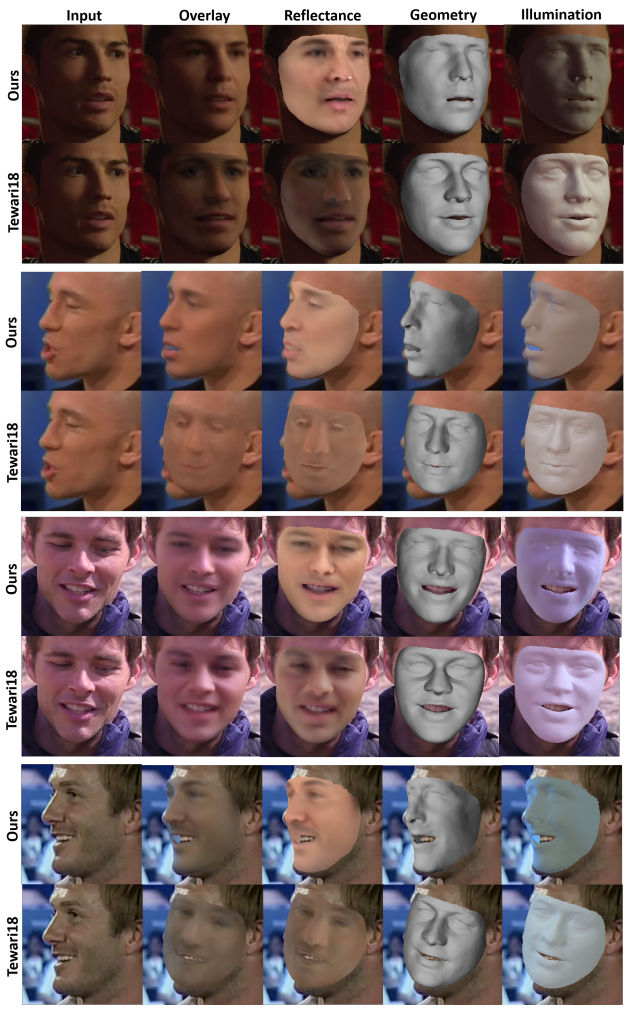


Figure 6. Comparison to Tewari et al. [8]. Multi-frame based training improves illumination estimation. Our approach also outperforms that of Tewari et al. when the face is largely occluded.

also successfully disentangles facial geometry, reflectance and scene illumination. Fig. 8 and 9 show that we obtain a fairly clean reflectance estimation, up to a small global scaling factor. Current state-of-the-art methods, on the contrary, only estimate facial texture that bakes in shading effects [11, 2]. We remark that our approach learns a reflectance model from scratch using only a colored template mesh, whereas the method of Booth et al. [2] require a 3DMM as initialization to learn a texture model, see Fig. 9.

**Quantitative evaluations** We quantitatively evaluate the photometric error of our approach on 1000 images of the CelebA dataset [4], see Fig. 11 and Tab. 4. We achieve lower errors when using larger models for shape and geometry. We also obtain lower errors compared to the 3DMM-based optimization approach presented in [3]. This demonstrates better generalization capabilities of our learned shape and appearance models to in-the-wild images, compared to the fixed face model [1] used by [3].

Fig. 12 shows the quantitative evaluation of our geometry reconstruction on 324 images of the BU-3DFE dataset [12]. Training on multiple frames consistently improves reconstruction quality. Multi-frame reconstruction with two

Table 4. Average photometric error (R,G,B $\in [0, 255]$) over 1000 images of the CelebA[4] dataset. Size refers to the number of vectors in our learned shape and appearance models. Larger models lead to lower errors. Our method outperforms [3] which reconstructs faces using an existing face model [1].*

| | Ours | | | | | [3] |
|---|---|---|---|---|---|---|
| Size | 0 | 10 | 50 | 125 | 500 | 80 |
| Mean | 32.54 | 23.15 | 21.03 | **20.66** | 20.82 | 21.95 |
| SD | 8.88 | 6.68 | 6.21 | 6.12 | 6.09 | 5.60 |

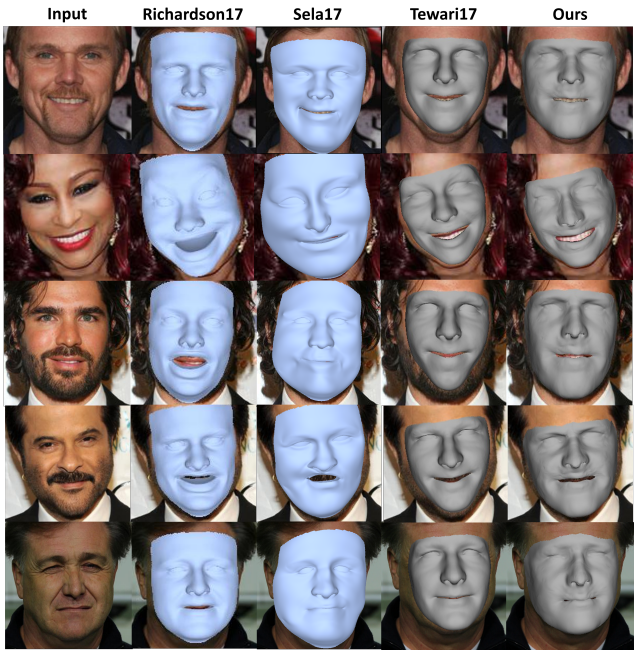


Figure 7. Comparison to [6, 7, 9]. These approaches are constrained by the (synthetic) training corpus and/or underlying 3D face model. Our optimal learned model produces more accurate results, since it is learned from a large corpus of real images.

images at test time also increases reconstruction quality compared to the monocular reconstruction case. We remark that our approach outperforms that of Tewari et al. [9, 8] on this dataset.

**Evaluation of different pooling techniques** We use average pooling for fusing features extracted from the multi-frame images. We evaluate the performance of multi-frame reconstruction using max pooling as well. With $M = 4$ at training and $M = 2$ at test time, both pooling techniques lead to very similar geometric reconstruction errors on BU-3DFE (difference is less than $0.004$mm). However, other pooling techniques could further improve our results.

## 4. Limitations

In this paper, we have proposed a multi-frame self-supervised deep learning approach that jointly learns a 3D face model (3D geometry and facial identity) and reconstructs 3D faces from in-the-wild videos. Although we have

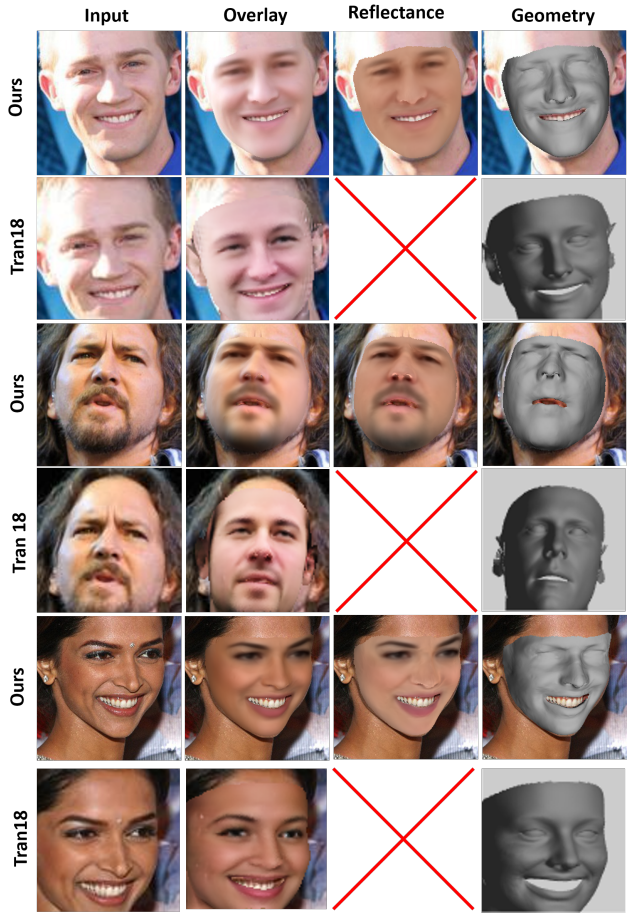


Figure 8. In contrast to Tran et al. [10], we estimate better geometry and separate reflectance from illumination. Note, the approach of Tran et al. does not disentangle reflectance and shading.

shown compelling results, our approach still has a few limitations that can be addressed in follow-up work, see Fig. 10.

Overall our approach can deal with large head poses quite well. Still, reconstructing extreme poses is a hard task in itself that challenges all face reconstruction techniques. Occlusions, e.g., by accessories or thick facial hair might adversely impact the reconstruction quality of our approach. Facial hair, such as beards are modeled in the reflectance channel, and thus are not reconstructed in a physically cor-

*There have been minor changes to the quantitative numbers reported in the paper, due to an error in the data loading script. The new numbers do not change any claims regarding comparisons to the state-of-the-art.

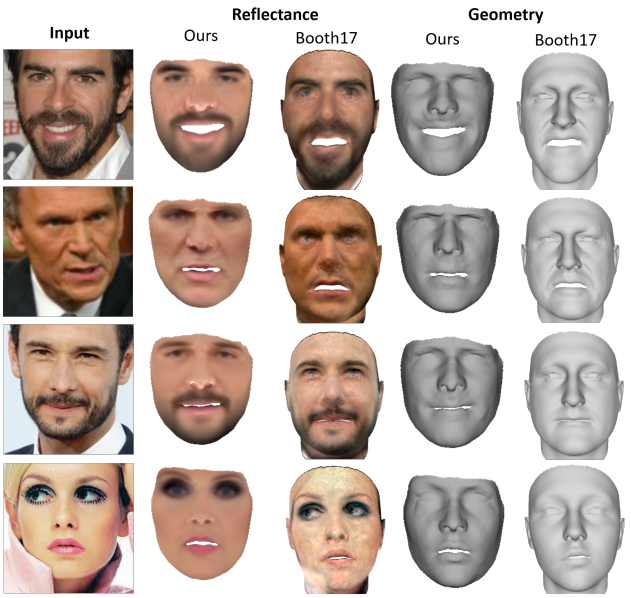


Figure 9. In contrast to the texture model of Booth et al. [2] that contains shading, our approach estimates a reflectance model.


Figure 10. Limitations of our approach. From top to bottom: Extreme illumination conditions, severe occlusions by accessories, thick facial hair, non-average facial shapes, scale ambiguity between illumination and reflectance, and extreme head poses.

rect manner. Even though our multi-frame supervision approach can obtain quite clean reflectance estimates that are free of shading, there is still a remaining global scale ambiguity between illumination and reflectance. As such, the global skin tone can not be reliably disentangled from the general ambient brightness of the illumination. Strong and colorful directional illumination outside the norm might also harm the estimation of 3D faces. Specular reflections and cast shadows are currently not modeled by our differentiable renderer, and thus they might slightly be baked into the reflectance channel. Non-standard facial shapes challenge our approach. We remark that all of these are difficult settings for almost any face reconstruction technique. Our approach already handles the aforementioned cases quite well by learning from in-the-wild videos without any sort of explicit 3D supervision.

## References


[1] V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *Proc. SIGGRAPH*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 1, 3, 4

[2] J. Booth, E. Antonakos, S. Ploumpis, G. Trigeorgis, Y. Panagakis, and S. Zafeiriou. 3d face morphable models "in-the-wild". In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3, 5

[3] P. Garrido, M. Zollhöfer, C. Wu, D. Bradley, P. Pérez, T. Beeler, and C. Theobalt. Corrective 3d reconstruction of lips from monocular video. *ACM Trans. Graph.*, 35(6):219:1–219:11, 2016. 3, 4, 6

[4] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 3, 4, 6

[5] E. Richardson, M. Sela, and R. Kimmel. 3D face reconstruction by learning from synthetic data. In *3DV*, 2016. 2

[6] E. Richardson, M. Sela, R. Or-El, and R. Kimmel. Learning detailed face reconstruction from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 4

[7] M. Sela, E. Richardson, and R. Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to-Image Translation. In *ICCV*, 2017. 2, 4

[8] A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4

[9] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Perez, and T. Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, 2017. 2, 4

[10] L. Tran and X. Liu. Nonlinear 3d face morphable model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Salt Lake City, UT, June 2018. 4

[11] L. Tran and X. Liu. On learning 3d face morphable model from in-the-wild images. arXiv:1808.09560, 2018. 3

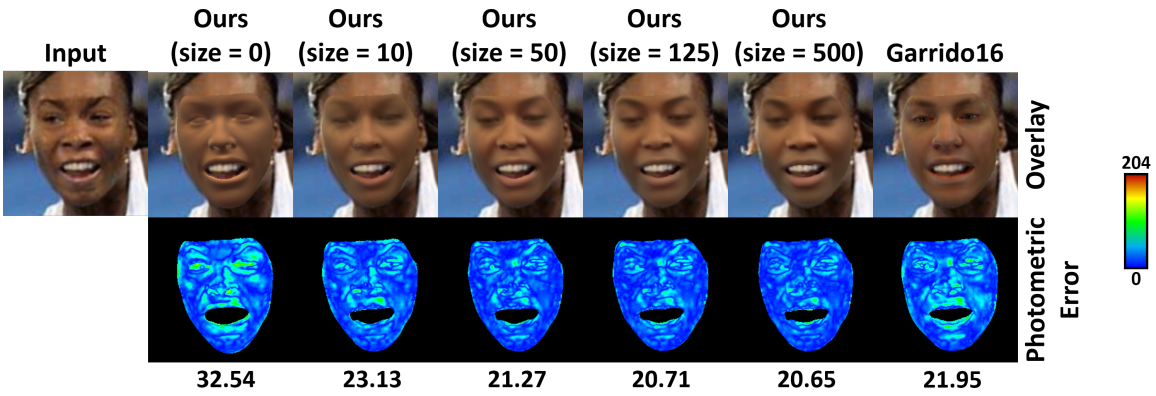


Figure 11. Quantitative evaluation of photometric error on the CelebA [4] dataset. Size is the number of learnable vectors in our shape and appearance models. Our method outperforms [3] which uses an existing model for reconstruction. The numbers are the average photometric errors (R,G,B $\in [0, 255]$) over 1000 images of the CelebA[4] dataset.

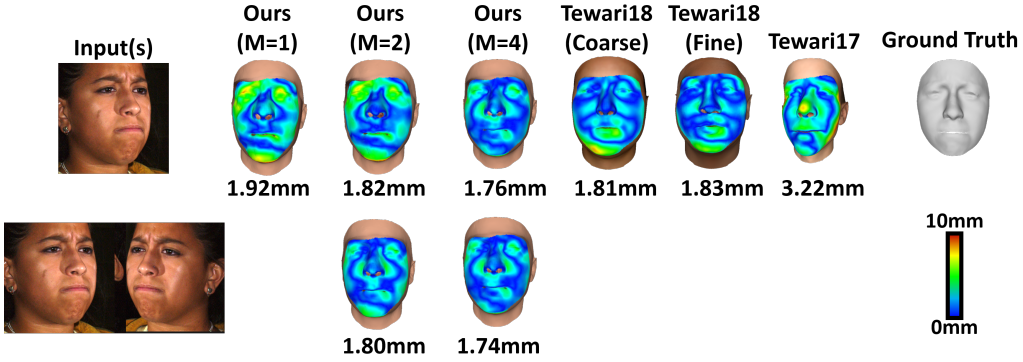


Figure 12. Quantitative evaluation on the BU-3DFE [12] dataset. The numbers are the geometric reconstruction errors averaged over 324 meshes. $M$ is the size of the multi-frame images used at training time. Multi-frame inputs at training and at testing time help in obtaining better reconstructions.

[12] L. Yin, X. Wei, Y. Sun, J. 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, 6