3D Facial Expressions through Analysis-by-Neural-Synthesis
Supplementary Material

George Retsinas[1†] Panagiotis P. Filntisis[1†] Radek Daněček[3] Victoria F. Abrevaya[3]
Anastasios Roussos[4] Timo Bolkart[3*] Petros Maragos[1,2]

[1]Institute of Robotics, Athena Research Center, 15125 Maroussi, Greece
[2]School of Electrical & Computer Engineering, National Technical University of Athens, Greece
[3]MPI for Intelligent Systems, Tübingen, Germany
[4]Institute of Computer Science (ICS), Foundation for Research & Technology - Hellas (FORTH), Greece

This supplementary material provides additional details and results for SMIRK. Section 1 describes the architectural choices and training details. In Section 2, we provide further quantitative evaluations, and Section 3 presents an extended set of ablation studies to better understand the impact of various components and design decisions. Finally, in Section 4, we discuss the limitations of SMIRK and explore potential future research directions, and Section 5 showcases more qualitative results.

1. Implementation Details

We describe here the implementation details of various subcomponents of the proposed method. For more information we refer to our method's source code and demo video: https://georgeretsi.github.io/smirk/.

1.1. Image-to-Image Translator

One important component in the proposed pipeline is the *Image-to-Image Translator*, which relies on UNet architecture [17]. Figure 1 depicts this module and all its subcomponents. In more detail, our implementation comprises the typical encoder and decoder convolutional parts, connected with shortcut paths, as shown in Fig. 1. Additionally, between the encoder and the decoder, we used a set of residual layers to further process the encoder output. The core feature of this module is the shortcut connections, either as residual connections or as UNet connections, that allow the gradients to be easily propagated through the entire network. As mentioned before, this image-to-image translation operation should be an appearance-first model, since the geometry of the face is given through the rendered 3D face and the main functionality of the translator resides in inpainting the missing texture. We validate the importance of shortcut connections in the ablation study of Sec. 3.3.

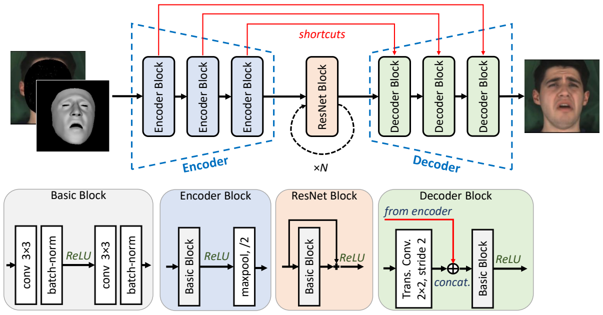

Figure 1. **Architectural Overview of the Image-to-Image Translator.** The encoder, which consists of 3 encoder blocks, downscales (/8) the initial input into a feature tensor map of size $H/8 \times W/8 \times 512$. This feature map is further processed through a set of residual blocks. The image is then reconstructed through the decoder, which consists of 3 decoder blocks. These decoder blocks upscale the feature maps using transposed convolutions, concatenate the resulting feature map with the respective map from the encoder phase using shortcut connections, and process the output with typical convolution operations (Basic Block).

1.2. Transfer Pixels in Cycle Path

One simple, yet effective, component of the augmented cycle path is the *transfer pixel* operation. In the cycle path we have a new tweaked expression and thus the facial points that we have selected from the initial image correspond to translated points in the new augmented image. If we keep the pixel locations as they are, from the initial image, inconsistencies will arise. For example, a pixel that corresponds to the lips in the initial image may correspond to the mouth interior in the tweaked expression.

[*] Now at Google.
[†] Equal contributions.

Given an initial expression and the new imposed expression, we know the difference between the two corresponding face geometries. In other words, if we select a pixel that corresponds to a facial point at the initial image, we can calculate the displacement vector that maps it to the new pixel location of the same facial point at the image with the tweaked expression. In this way, we can sample facial locations that are consistent. This observation is the core of this functionality, where we sample some pixels based on the facial geometry of the initial predicted expression, we displace the pixel positions according to the new expression and we assign them the RGB values coming from the initial pixel locations. Formally, given a sparse set of selected pixels with positions $\{x_i\}$ on the initial image I, we create an augmented "guidance" image I_{aug}, that samples the interior of the new face, using the displacement vectors $\{d_i\}$ as $I_{aug}(\lfloor x_i + d_i \rceil) = I(x_i)$ for each (x_i, d_i) pair. Note that image values are RGB triplets.

1.3. Identity Loss

Preliminary versions of the SMIRK framework did not include the *transfer pixels* operation. Thus we used pixels of the initial un-tweaked image as guidance in the cycle path of different expressions. This introduced an inconsistency between reconstruction and cycle path and cycle image reconstruction were non-realistic, following only the rendered expression. To address this we used an off-the-self perceptual identity loss, implemented via a Resnet50 model pretrained on the VGG-Face2 dataset [4, 7].

Nonetheless, for the final SMIRK version, where we use the transfer pixels option, the aforementioned issue is minimized. Instead, we use a *structural* identity loss. As discussed in the main manuscript, this loss uses the frozen shape encoder $E_{\boldsymbol{\beta}}(I)$ to enforce a structural shape consistency by minimizing the L_2 distance between the predicted shape and the original shape. This loss acts only on the image-to-image translator T and tries to generate accurate image reconstruction by promoting decoupling of the shape/expression parameters.

1.4. Template Injection

In order to acquire templates (i.e., expression parameters) that correspond to specific, rarely-encountered expressions, we have performed direct iterative parameter fitting on the FaMoS [2] dataset. More specifically, we fitted pose and expression parameters of FLAME to the following sequences of the dataset from 70 random subjects, using a sampling stride of 10: lips back, rolling lips, mouth side, kissing, high smile, mouth up, mouth middle, mouth down, blow cheeks, cheeks in, jaw, lips up. To ensure accurate results we used the corresponding neutral template provided for each subject, instead of optimizing the identity parameters. For parameter fitting we used the official tensorflow implementation [12] provided by the authors of FLAME. We present examples of these expression templates using the mean FLAME identity in Figure 2.

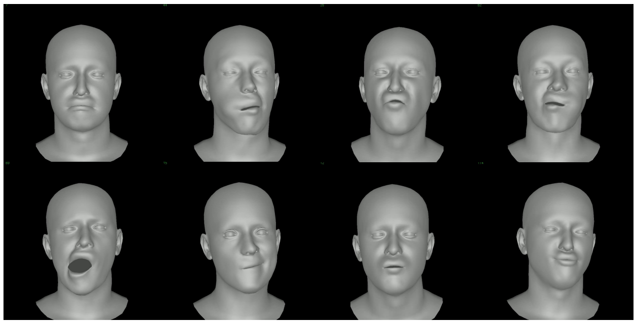

Figure 2. Examples of expression templates used in the cycle path.

1.5. Model Sizes

In this work we aimed for a more lightweight encoder, and hence used MobileNetv3 [9] backbones. Table 1 reports the number of parameters for SMIRK and the other considered methods. As we can see, SMIRK is 14 times smaller than EMOCA/EMOCAv2, and 7 times smaller than other state-of-the-art methods. These results further strengthen the superiority of SMIRK, since the considered encoder is of limited capacity.

	SMIRK	DECA	EMOCAv2	FOCUS	Deep3d
# Params	**3.6M**	26.8M	51.4M	25.5M	24.0M

Table 1. Number of parameters in SMIRK and other SOTA models. SMIRK is 14 times smaller than EMOCA and 7 times smaller than the other methods.

1.6. Training details

Pretraining: Before training the expression encoder of SMIRK we pretrain all encoders using only landmark losses. During this step a shape regularizer is also added to impose identity shaping with respect to a pre-trained network (MICA [23]). The pretraining phase is done for 60,000 iterations using Adam with a learning rate of $5e-4$.
Face Rendering: FLAME is a full head model which includes ears, eyeballs, neck, and scalp in the facial mesh. However, in our work we only render the *expressive* part of the 3D model, which is the face. Images of this rendering can be seen in the pipeline figures in the main paper.
Training: We use the following datasets for training: FFHQ [10], CelebA [13], LRS3 [1], and MEAD [19]. Since LRS3 and MEAD are video datasets, we randomly sample images from each video during training. We train using a batch size of 32, where each batch consists of 50% images from FFHQ and CelebA to promote in-the-wild reconstruction, 40% images from MEAD to promote the emotional ex-

	mean ↓	median ↓	max ↓
DECA	1.40	1.12	6.8
EMOCAv1	1.45	1.19	6.83
EMOCAv2	1.43	1.15	6.78
SMIRK	**1.28**	**1.05**	**5.98**

Table 2. Per-vertex 3D reconstruction errors (mm) on MultiFace[20]. SMIRK outperforms other FLAME-based methods.

pressions seen in this dataset, and 10% images from LRS3, to promote diverse mouth formations during speech. The weights of the losses used for training are $\mathcal{L}_{cycle} = 10$, $\mathcal{L}_{lmk} = 100$, $\mathcal{L}_{vgg} = 10$, $\mathcal{L}_{photo} = 1$, $\mathcal{L}_{emo} = 1$, $\mathcal{L}_{reg} = 1e{-}3$. In the Augmented Expression Cycle Path we augment each predicted sample uniformly with one for each of the augmentations that were described in the main paper. During the core phase we train SMIRK for 250,000 iterations with a learning rate of $1e - 3$ and cosine-annealing, restarted at each epoch.

Landmarks: For the landmark loss, like EMOCAv2 [5], we use a combination of 92 predicted mediapipe landmarks for the interior of the face and 16 landmarks from FAN[3] for the face boundary.

2. Additional Quantitative Results

Although as we mentioned in the main text, geometric errors tend to not correlate well with human perception, we also present here the per-vertex errors on the MultiFace [20] datasets for all FLAME-based methods (which have the same topology). The MultiFace [20] v1 dataset consists of 3D scans captured in a multi-camera setup, where subjects where asked to perform various extreme facial expressions. To evaluate the per-vertex error we select the frontal camera subset and select the subjects whose face is fully shown in the image. We use the official test set ("EXP_ROM07_Facial_Expressions"), resulting in a total of 6,324 facial expressions across 5 subjects. In Table 2 we report the mean, median, and max of the ScanToMesh[18] distances between the scans and the predicted mesh surfaces from all FLAME-based methods. Note that the max per-vertex error has been previously reported to correlate better with perceptual quality, compared to the mean that tends to mask inaccurate expressions [16, 22]. As we can see, SMIRK outperforms the other methods on all 3D-reconstruction metrics, and significantly reduces the maximum 3D reconstruction error. Figure 3 also shows qualitative comparisons where SMIRK captures significantly more faithfully extreme and asymmetric expressions.

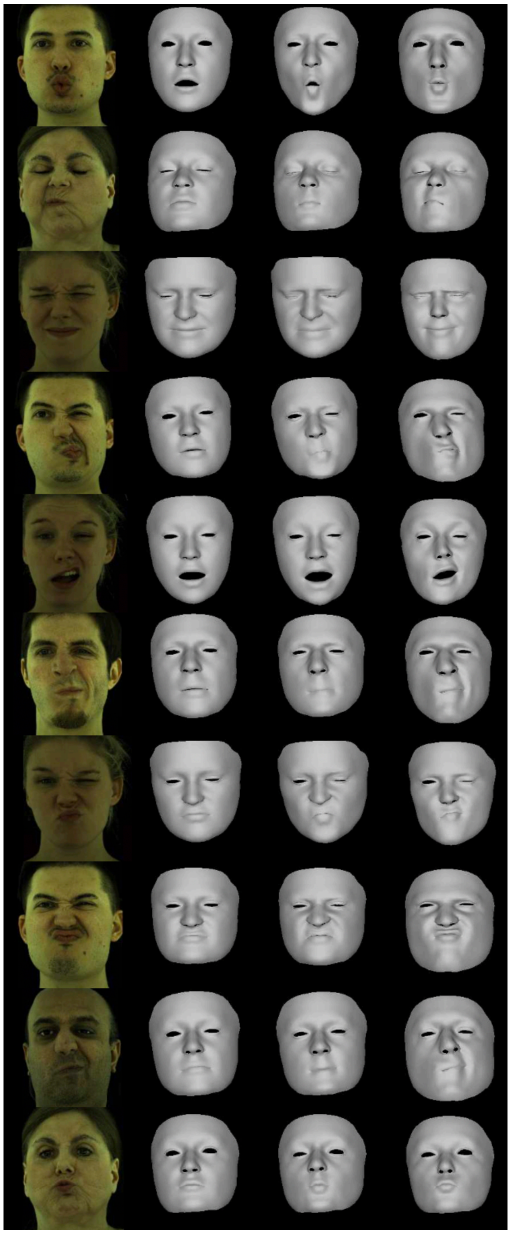

Figure 3. Qualitative comparison of FLAME-based methods on the Multiface dataset. From left to right: Input, DECA[7], EMOCAv2[5], SMIRK. SMIRK excels in capturing extreme and asymmetric expressions.

3. Additional Ablation Studies

In this section we explore the impact of several proposed architectural/training options.

3.1. Impact of Masking

The proposed masking process selects a small number of random pixels inside the face to provide useful texture-related information for the reconstruction of the image. We

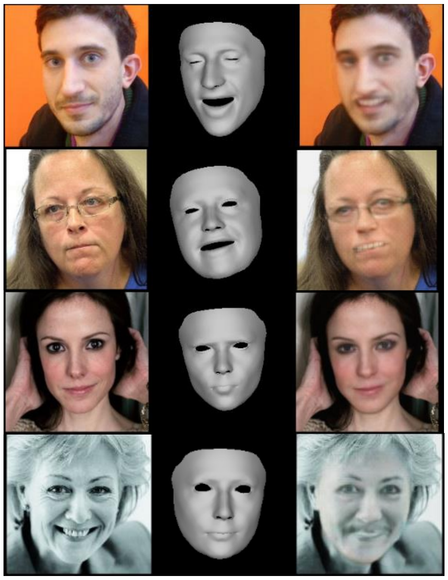

Figure 4. **Masking with higher percentage of retained pixels.** Left: initial image, Middle: target manipulated expression, Right: reconstructed image. The ratio of pixels to be retained was set to 5% instead of the default 1%. We observe that the mouth and eyelid opening/closing cannot be captured adequately, and the emotion is not transferred from the manipulated expression.

have mentioned that a very small number of pixels is retained, i.e. only 1%, since using a higher percentage usually leads to non-realistic inpainting actions. Such cases are depicted in Fig. 4, where 5% of the pixels are retained. As we can see, the image reconstruction step struggles to capture different expressions since it relies too much on the selected pixels, with mouth and eyes opening/closing being a major problem. Moreover, emotions cannot be correctly manipulated, as the reconstructed image retains the emotion of the initial image (see e.g. 3rd row of Fig. 4).

3.2. Impact of Cycle Path

Here, we perform ablation studies regarding the accuracy of SMIRK with and without the extra augmented expression cycle path, which enables the encoder to see more variations in expressions and further promotes consistency.

Image reconstruction First, using the protocol in Section 4.2 of the main paper, we train from scratch a UNet image-to-image translator for both the encoders with and without the cycle path. We then calculate the reconstruction losses (L_1 and VGG losses) on the test set of AffectNet. These results can be seen in the first two columns of Table 4. As we can see, both encoders (with and w/o cycle path) have a very close performance w.r.t. reconstruction metrics, indicating a good correspondence, in average, between the rendered 3D face and the initial image for both alternatives.

	mean ↓	median ↓	max ↓
no cycle path	1.43	1.16	6.69
no-injection	1.32	1.07	6.08
no-permutation	1.33	1.07	6.09
no-zeroing	1.34	1.08	6.12
no-random	1.33	1.07	6.12
all augments	**1.32**	**1.07**	**6.02**

Table 3. Ablation study on the effect of different cycle augmentations on the MultiFace dataset (per-vertex 3D reconstruction errors in mm).

	L_1 Loss ↓	VGG Loss ↓	vert L_1 ↓	vert abs std ↓
w/o cycle path	0.096	**0.758**	0.0130	0.0068
w/ cycle path	**0.095**	0.762	**0.0128**	**0.0044**

Table 4. **Image reconstruction performance** with and without the cycle loss, evaluated on the AffectNet test set [15]. First two columns correspond to the reconstruction metrics, whilst the latter two measure the capability of the generated images to capture changes in expression.

Capturing small variations However, these metrics cannot evaluate the capability of the network to capture small variations in expression. To do so, we devised a more in-depth ablation study that highlights the adaptability (w.r.t. to expression changes) of a trained translator with and without the cycle path. Starting from the inferred FLAME parameters on an input test image (from the test set of AffectNet), we apply N different (minor) augmentations in the expression parameters within a batch, including jaw and eyelids. Then, we use the image-to-image translator to generate a variant of the input face with the new expression and we re-apply the trained encoder to obtain the re-estimated expression parameters, akin to the cycle operation. Finally, we calculate:

- the L_1 norm, dubbed as *vert* L_1, between the 3D vertices corresponding to the initial tweaked set of expression parameters, that was used to generate the photorealistic copy, and the 3D vertices corresponding to predicted set of expression parameters. We use the comparison on the vertices space to avoid penalizing possible ambiguities in the expression space that the alternative without cycle loss cannot easily discern.
- the absolute difference between the standard deviation of the N different copies of each input face, dubbed as *vert abs std*. Again, we calculate this metric between the corresponding vertices. This metric indicates how well the encoder can identify minor changes in expressions.

The aforementioned metrics can be found in the last two columns of Table 4. As can be seen, using the cycle path results in similar L_1 performance with the non-cycle option (the cycle variant is marginally better), but preserves considerably better the standard deviation between the different image copies. The latter is a strong indicator that training

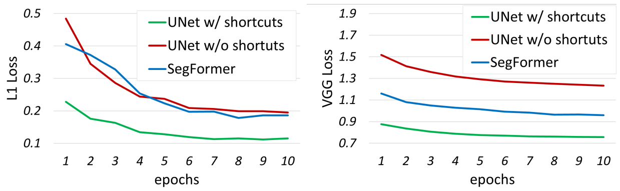

Figure 5. **Image reconstruction performance for different Translators,** using L1 loss (left) and VGG loss (right).

with the proposed cycle path helps retaining the variability of the expression parameter space through the translator.

Note that the encoder trained with the cycle path option has seen reconstructed images and used them as input, whilst the encoder train without the cycle path has not, Thus, to ensure that these improvements using the cycle path are not fictitious due to a possible distribution shift between the generated images of the two alternatives, we re-run the above experiment without tweaking the original expression. Thus, both encoders are used on the non-altered reconstructed images as sanity check, in essence validating if the generated images in both cases are realistic enough and close to the initial domain. In this case both encoders performed equally well (0.0129 without cycle path and 0.0128 with cycle path), which shows that no notable domain shift, capable of favoring the one alternative over the other, is evident.

Per-vertex reconstruction error Finally, we also assess the impact of the cycle path and the different augmentations in terms of 3D per-vertex reconstruction error. To do this we train separate models for 20 epochs on FFHQ. Results can be found in 3, for the MultiFace dataset. As we can see, best results occur with all augmentations combined, while removing individual augmentations leads to decreased results. Removing the cycle path completely, considerably drops performance.

3.3. Impact of Translator's Architecture

One critical property of the Translator, under the proposed framework, is the "uninterrupted" gradient flow. As we have already described, this is achieved through the shortcut connections of the proposed architecture (as shown in Fig. 1). To validate the importance of these shortcuts connections, we simulate the same architecture (exact same number of parameters) without shortcuts (neither UNet nor residual shortcuts). We also consider the transformer-based SegFormer architecture [21] as an alternative. The UNet variants have $\sim$ 30M trainable parameters, while the SegFormer has $\sim$ 85M parameters.

We trained these three architectural variants with the proposed framework for 10 epochs. The evaluation protocol is the same as in Sec 3.2. The progress of L1 and VGG losses through these 10 epochs is depicted in Figure 5. We can observe that the default option with the shortcut connections has a fast convergence to meaningful reconstructions, letting the framework to focus on discovering subtle expression details. The other alternatives struggle in the first epochs to adapt to the image reconstruction task, which may have a negative impact on the 3D prediction step.

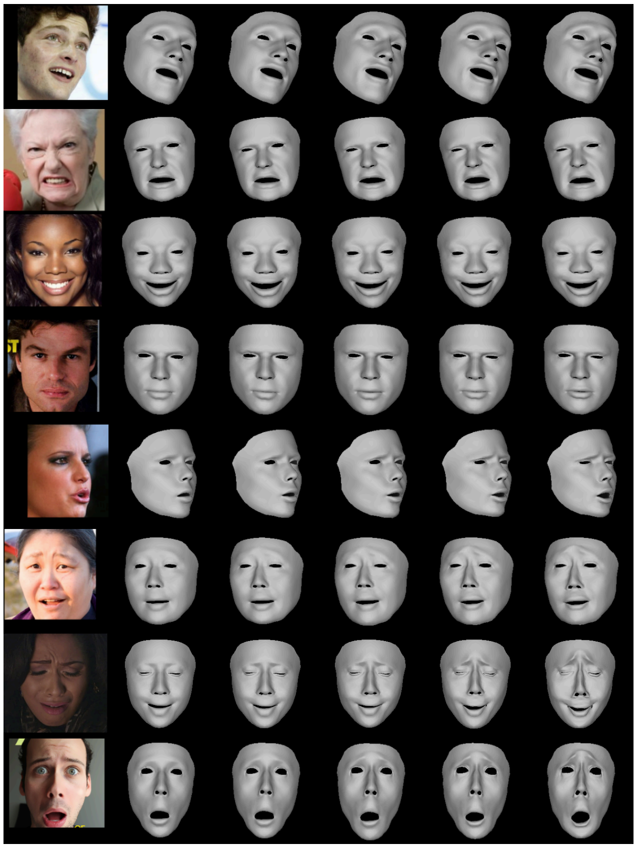

Figure 6. Image results on the effect of emotion loss weight. From left to right, $\mathcal{L}_{emo} = 0, 1, 2, 5, 10$. We see that in certain cases, higher emotion losses can lead to exaggerated expressions and artifacts.

3.4. Impact of Emotion Loss

One of the advantages of the proposed approach is the direct comparison between the reconstructed image and the input image via perceptual losses, without any domain gap involved. In this work, we considered an extra emotion loss, following EMOCA [5]. The goal is straightforward: assist the encoder to better capture emotion-related expressions.

One can tune the contribution of this auxiliary loss through its respective weight, used to calculate the overall loss. Using very small values has minor to no impact, while large values cause over-exaggerations of the requested emotions, leading to visually unfaithful 3D reconstructions, as was the case in EMOCA [5]. The emotion weight was set to 1, as the default option, after visual inspection for possible

emotion weight	V-PCC ↑	V-CCC ↑	V-RMSE ↓	V-SAGR ↑	A-PCC ↑	A-CCC ↑	A-RMSE ↓	A-SAGR ↑	E-ACC ↑
0 (w/o emot. loss)	0.72	0.71	0.35	0.79	0.62	0.60	0.32	0.79	0.62
1 (default)	0.72	0.72	0.35	0.79	0.64	0.61	0.31	0.79	0.64
2	0.73	0.71	0.34	0.80	0.62	0.60	0.32	0.80	0.62
5	0.75	0.74	0.33	0.81	0.65	0.63	0.31	0.77	0.65
10	0.74	0.72	0.33	0.79	0.64	0.63	0.32	0.76	0.63

Table 5. **Emotion recognition results** for different emotion weights.

method	neutral↑	happy↑	sad↑	surprise↑	fear↑	disgust↑	anger↑	contempt↑	avg. (macro)↑
Deep3D	0.62	0.81	0.62	0.60	0.72	0.59	0.60	0.53	0.63
DECA	0.45	0.76	0.51	0.50	0.69	0.64	0.59	0.54	0.58
FOCUS	0.50	0.75	0.47	0.65	0.51	0.61	0.56	0.58	0.58
EMOCA v1	0.50	0.79	0.69	0.67	0.69	0.76	0.70	0.66	0.68
EMOCA v2	0.52	0.77	0.58	0.66	0.66	0.73	0.73	0.68	0.67
SMIRK w/o em. loss	0.55	0.72	0.54	0.65	0.58	0.62	0.58	0.70	0.62
SMIRK $\mathcal{L}_{emo} = 1$	0.56	0.79	0.60	0.64	0.57	0.70	0.61	0.59	0.64
SMIRK $\mathcal{L}_{emo} = 2$	0.44	0.74	0.59	0.65	0.61	0.59	0.66	0.69	0.62
SMIRK $\mathcal{L}_{emo} = 5$	0.60	0.77	0.63	0.64	0.58	0.66	0.59	0.70	0.65
SMIRK $\mathcal{L}_{emo} = 10$	0.47	0.75	0.56	0.67	0.64	0.62	0.63	0.74	0.64

Table 6. **Accuracy per emotion** for all methods and average (macro).

expression over-exaggerations.

Using the protocol of [5], and complementary to the results in Section 4.2 of the main paper we show the effect of different emotion weight in Table 5. In addition, more in-depth exploration of the emotion recognition performance is given in Table 6, where we also report per-emotion accuracy, along with the average across emotions, for different emotion weights, as well as the considered SOTA methods.

We observe that different emotion weights can result in different and non-canonical pertubations in the results, e.g., for emotion loss weight 1 the accuracy for contempt drops drastically w.r.t. using no emotion, while a similar effect occurs when increase the weight from 5 to 10 for sadness. We also see that the trained MLPs tend to confuse the negative emotion (fear, disgust, anger), and more succesfully predict happiness. Overall, this behavior of the emotion recognition results could be attributed to a possible sensitivity of the trained MLPs combined with the ambiguous nature of emotion classification. In Figure 6 we also show qualitative examples on the effect of emotion loss on images from the AffectNet dataset. We can see that in many cases (rows 1 - 4) the effect of emotion loss weighting is smaller, however for certain emotions san as sadness, the results tend to get very exaggerated as the emotion loss increases. This can result in serious artifacts with higher emotion losses (see last 2 rows).

In Figure 7 we also show some qualitative examples comparing SMIRK, against EMOCAv1 and EMOCAv2. As it can be seen, EMOCAv1 which achieves the highest emotion recognition accuracy under this protocol tends to significantly exaggerate the observed emotion. On the other hand, EMOCAv2 often lacks the visual consistency with the original face. This could be attributed to the domain mismatch in the emotion recognition loss in EMOCA, since a textured rendered face with albedo is compared with the original image.

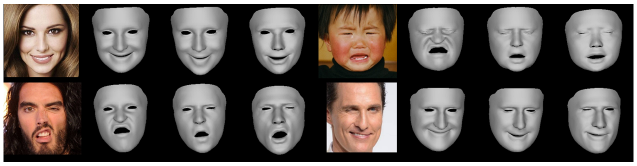

Figure 7. **3D reconstruction of emotions**. From left to right: input image, EMOCA v1, EMOCA v2, SMIRK. EMOCA v1 tends to exaggerate emotions, hence the highest score in emotion recognition. EMOCA v2, on the other hand, often lacks visual consistency with the original face, possibly due to a domain mismatch in the employed emotion recognition loss.

3.5. Expression Pretraining Ablation

We also evaluate the proposed pipeline, when the expression encoder is not initialized by the pre-trained network and show results in terms of 3D reconstruction in Table 7 and qualitative in Figure 8. As we can see, training the expression encoder from scratch achieves similar and comparable results compared to using a pretrained expression encoder on landmarks only.

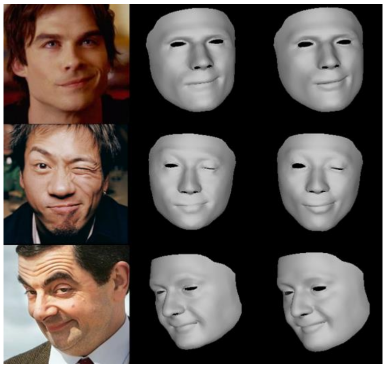

Figure 8. SMIRK with (middle column) and without (right column) pretraining the expression encoder achieves comparable results.

	mean ↓	median ↓	max ↓
SMIRK w exp. pretrain	1.28	1.05	5.98
SMIRK w/o exp. pretrain	1.31	1.08	6.07

Table 7. Ablation study on the effect of pretraining the expression encoder on the MultiFace dataset (per-vertex 3D reconstruction error in mm).

4. Limitations & Future Directions

Despite the effectiveness of the proposed method, there are specific limitations to be addressed, each one of them constituting a potential future direction:

- *Occlusions, Extreme Poses, and Challenging Lighting Conditions:* The majority of the datasets used in the proposed method have limited occluded cases and extreme poses. This makes the method sensitive to occlusions, as it tends to assume more intense expressions where a part is missing, rather than extrapolating from the existing information and retaining a more ”average” expression for the missing parts. Additionally, the method can produce degraded results under cases with very limited lighting, as demonstrated in Figure 9. Nonetheless, addressing such cases was not within the scope of this work.
- *Temporal Consistency:* the proposed framework has been trained on single images and the temporal aspect is not explored. Smooth temporal transition and consistency can be imposed through external losses for video input. Towards this concept, one could extend the set of perceptual losses by adding a lip reading term, as in [8].
- *Extension to Shape/Identity Parameters:* The present work focuses on estimating expression parameters, but the overall concept of learning through Analysis-by-Neural-Synthesis can be straightforwardly extended to estimate pose or identity parameters. Nonetheless, preliminary experiments showed that we cannot successfully optimize these parameters all-together, without sacrificing performance. Changing pose and shape each iteration affects the expression performance, not letting the expression parameters to capture finer subtle expressions due to “jittering” effects of continuously changing pose/identity. Nonetheless, given a good pose and expression estimation, one could fine-tune the shape parameters etc. Of course, optimizing shape also requires an extra set of regularization losses (e.g., shape consistency between different pictures of the same person).

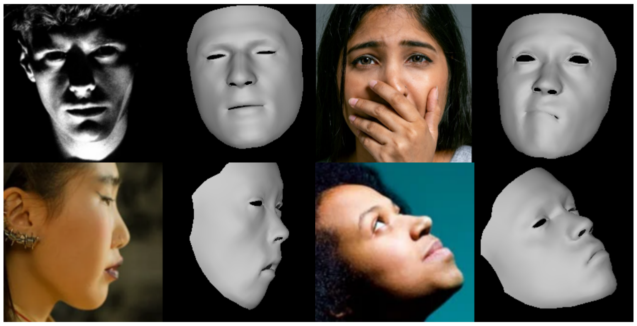

Figure 9. Examples where the SMIRK produces degraded results due to occlusions, extreme poses, and challenging lighting conditions.

5. Additional Qualitative Results

To further understand the effectiveness of SMIRK, we present a large set of visual examples in Figure 10, where our method is compared against other state-of-the-art approaches.

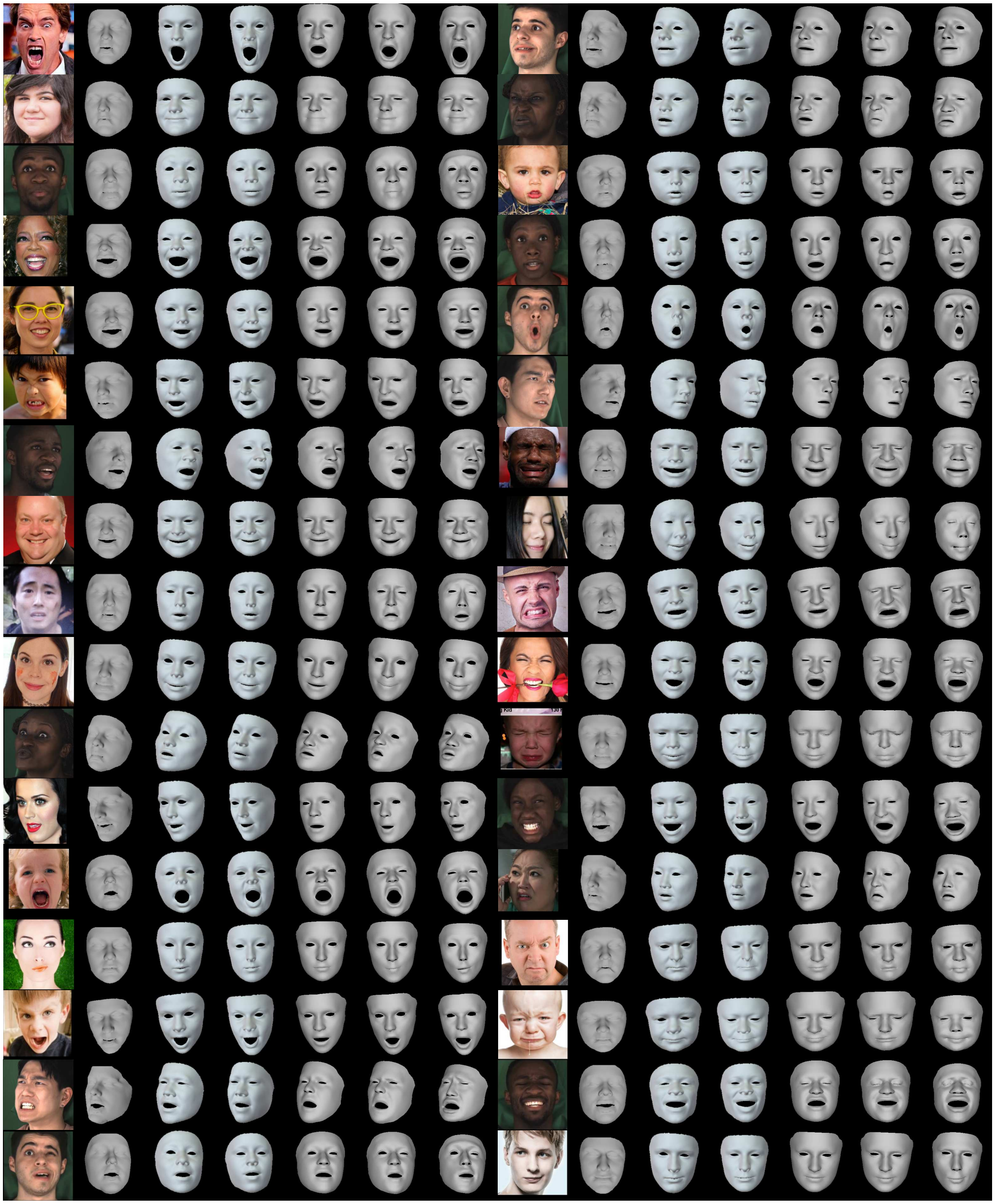

Figure 10. More qualitative results and visual comparisons of 3D face reconstruction from our method and four others. From left to right: Input, LeMoMo[14] (method results provided by the authors), Deep3DFaceRecon([6], FOCUS[11], DECA[7], EMOCAv2[5], and SMIRK. Please zoom in for details. Video results can also be found in the supplementary video.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Lrs3-ted: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2

[2] Timo Bolkart, Tianye Li, and Michael J Black. Instant multi-view head capture through learnable registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 768–779, 2023. 2

[3] Adrian Bulat and Georgios Tzimiropoulos. 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*, pages 1021–1030, 2017. 3

[4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2

[5] Radek Daněček, Michael J Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20311–20322, 2022. 3, 5, 6, 8

[6] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. Accurate 3D face reconstruction with weakly-supervised learning: From single image to image set. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, pages 285–295, 2019. 8

[7] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *Transactions on Graphics, (Proc. SIGGRAPH)*, 40(4):1–13, 2021. 2, 3, 8

[8] Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantoniou, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. SPECTRE: Visual speech-informed perceptual 3D facial expression reconstruction from videos. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*, pages 5745–5755, 2023. 7

[9] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 2

[10] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[11] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. To fit or not to fit: Model-based face reconstruction and occlusion segmentation from weak supervision. *CoRR*, abs/2106.09614, 2021. 8

[12] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. TF-FLAME. https://github.com/TimoBolkart/TF_FLAME, 2021. 2

[13] 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)*, 2015. 2

[14] B.R. Mallikarjun, Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3d morphable face models from images and videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3361–3371, 2021. 8

[15] Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017. 4

[16] Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182, 2021. 3

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, pages 234–241. Springer, 2015. 1

[18] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3d supervision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[19] Kaisiyuan Wang, Qianyi Wu, Linsen Song, Zhuoqian Yang, Wayne Wu, Chen Qian, Ran He, Yu Qiao, and Chen Change Loy. Mead: A large-scale audio-visual dataset for emotional talking-face generation. In *ECCV*, 2020. 2

[20] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Alexander Hypes, et al. Multiface: A dataset for neural face rendering. in arxiv, 2022. 3

[21] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 5

[22] Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, and Tien-Tsin Wong. Codetalker: Speech-driven 3d facial animation with discrete motion prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12780–12790, 2023. 3

[23] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pages 250–269, 2022. 2