Electromyography-Informed Facial Expression Reconstruction for Physiological-Based Synthesis and Analysis

Supplementary Material

Table of Contents

A. Implementation and Model Details

We provide an overview of the model architectures and experimental setups used in EIFER to facilitate re-implementation. This, combined with the publicly available source code[1], allows for a deeper understanding of EIFER's inner workings and suggests that the model architecture has a minor impact on the overall training pipeline.

EIFER is composed of three primary model components, which are duplicated for both the C^{SN} and C^{NS} cycles. Notably, during the evaluation of EMG2Exp and Exp2EMG, the C^{SN} cycle plays a crucial role. However, it is essential to recognize that the C^{SN} cycle cannot be trained in isolation from the other component, as the two cycles are interconnected and interdependent.

[1]Project page: https://eifer-mam.github.io

All models are implemented in PyTorch [37], and we utilize PyTorch3D [39] for rendering the FLAME [28] mesh to disentangle facial geometry from appearance.

A.1. Encoder and Face Model

We adopt the triple encoder structure from SMIRK [40] and utilize MobileNetV3 as the backbone network. This allows us to initialize EIFER with pre-trained SMIRK models, providing several benefits.

Firstly, the pre-trained models are assumed to be robust to rough alignment without facial landmarks, as demonstrated in the ablation studies of [40]. Secondly, we assume accurate facial feature extraction for non-sEMG occluded faces, enabling the other encoder to mimic the correct one under occlusion. Lastly, this initialization ensures comparability with existing SMIRK results, as the updated model parameters are robust to sEMG occlusion. The model architecture is illustrated in Figure 1.

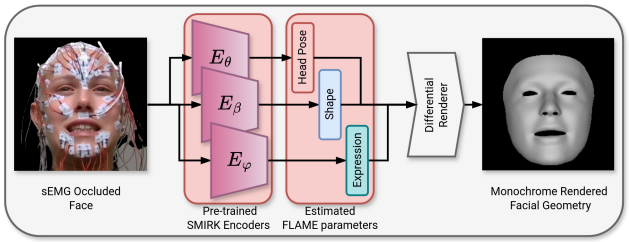


Figure 1. **EIFER Encoder Architecture** We utilize the tripe encoder setup of SMIRK [40] to predict the FLAME parameters [28]. Therefore, each sub-encoder can be used independently of the given task. In our case, EIFER updates the pre-trained models to handle sEMG occlusion.

We employ the intermediate FLAME [28] face representation, comprising 300 shape and 50 expression components to utilize pre-trained weights. Additionally, we include three components for jaw movement and two blendshapes for the eyelids [54]. The position sub-encoder estimates the head rotation and position, modeled by camera parameters.

We use these shape and expression parameters to construct the 3D FLAME mesh. A differential renderer [39] then generates a monochromatic render of the frontal face view. This rendered face contains essential facial geometry information following the same denomination as in [26, 40]. The generator model must restore the face correctly from this rendered face.

A.2. Generator Model

The generator model aims to reconstruct the input face faithfully. Unlike traditional rendering approaches [7, 8, 14, 26], we employ implicit neural rendering [2, 3, 40] for its robustness.

To disentangle facial geometry and appearance, we use image-to-image translation techniques. However, the input face image contains both geometry and appearance information. To address this, we use the rendered face image, computed by the encoder networks, as the primary driver. Additionally, we forward random pixel information from the input face to the generator, similar to [40], to recreate skin texture and lighting conditions.

The generator models take geometry and random appearance pixels as input, effectively functioning as an image-to-image translation network or style transfer model. Unlike traditional rendering approaches that rely on an appearance model [8, 17, 26, 38], we are not constrained by explicit assumptions, allowing us to adapt the generator models to our specific requirements.

To train the generator to ignore sEMG electrodes, we employ an unpaired reference image with a different expression and a discriminator. This setup has two benefits: (1) the model learns to ignore pixels describing sEMG electrodes, and (2) the generated faces must be photorealistic to convince the discriminator, eliminating the need for additional perceptual losses.

However, this adversarial problem poses challenges, such as generative models creating incorrect features or hallucinating wrong expressions. We refer the reader to the main paper for details on regularization terms that address these issues.

Unlike recent works [21, 40, 53], we use a ResNet [20] as the backbone architecture for our generator models. Although this differs from the typical Unet architecture, it allows for a similar gradient flow.

We modify the architecture to replace Conv2DTranspose layers with a single Upsample and Conv2D layer, eliminating the pixelated output and checkerboard patterns in SMIRK (see visualization in the main paper). This improves the overall quality of the generated images.

We employ instance normalization as the primary activation function throughout the network [45], which enhances the reconstruction quality and information flow in the optimization problem. However, instance normalization requires a batch size of one to avoid mirroring the behavior of standard batch normalization [45].

We adopt the multi-phase approach outlined in the main paper to address this limitation, as parallel-trained models like EMG2Exp cannot converge with small batch size. This approach ensures stable training and convergence.

Our ResNet Generator, shown in Figure 2, consists of 9 residual blocks with a feature depth of 64, similar to the parameter amount of the original UNet in SMIRK [40].

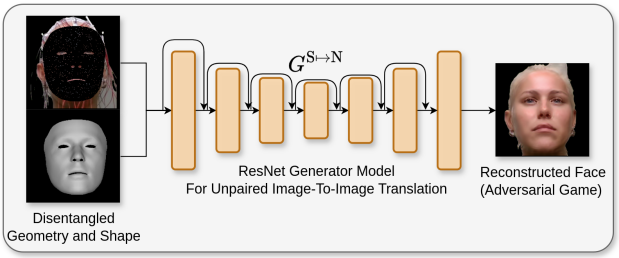


Figure 2. **EIFER Generator Architecture:** Our EIFER generator architecture is based on the ResNet [20] backbone, which serves as the neural generator for restoring faces. We incorporate skip connections to facilitate information flow, similar to the U-Net [41, 53] architecture employed in SMIRK [40]. However, we introduce two modifications: (1) we utilize instance normalization [45] to generate nuanced details, and (2) we replace the Conv2DTranspose layers with a combination of upsampling and convolutional layers to eliminate the checkerboard patterns.

A.3. Updated Masking Function

The masking function, originally proposed in [40], selects a random pixel to represent facial appearance. However, this function relies on computing facial landmarks in the input images to define a suitable sampling area. Unfortunately, this is not feasible under sEMG occlusion, as demonstrated in Figure 8.

We reformulate the sampling area based on the rendered FLAME face model to address this limitation. This is made possible by the sufficient pre-training of the encoder models, allowing us to tackle this complex problem without requiring a retrained sEMG occlusion-robust facial landmarking model.

As a result, EIFER implicitly becomes a robust facial landmarking tool under occlusion, as shown in the ablation studies. We illustrate the information flow of the updated masking function in Figure 3.

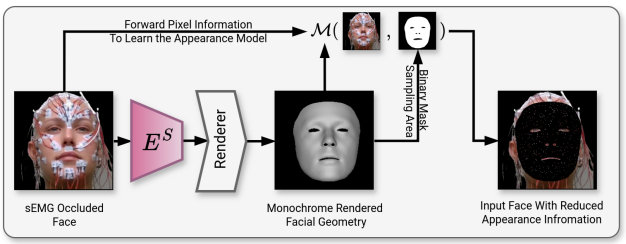


Figure 3. **Update Information Flow For the Pixel Masking:** As we cannot rely on the facial landmarks convex hull as a sampling area, we utilize the monochrome rendered facial geometry instead for the masking function $\mathcal{M}(\cdot)$ [40]. This has the advantage that we can utilize the learned alignment capabilities of SMIRK [40]. The selected sampling area covers the facial area well.

A.4. Discriminators

We employ a simple yet effective discriminator model inspired by previous works [2, 18, 53], distinguishing between generated faces and their unpaired reference images. Specifically, we compare generated faces with removed and applied sEMG electrodes, ensuring that the generator produces realistic faces consistent with the input data.

To train the generator, we utilize the least square GAN loss [33], which encourages the generator to produce more realistic faces by enforcing generation near the decision boundary. This loss function helps to stabilize the training process and improve the overall quality of the generated faces. Consequently, the problem of distinguishing between real and fake faces is now reduced to a two-class decision problem. Our discriminator network consists of a 3-layer convolutional neural network with two output neurons, which classify input images as real or fake.

A.5. Multi-Stage Training

We adopt a multi-stage training approach for the two encoder-generator pairs and the two-phase training protocol to overcome the batch size limitation. This approach is critical due to the challenging nature of our problem, where facial features are obstructed by electrodes, making expression extraction difficult.

Inspired by previous works [2, 3], we employ a two-stage training strategy. In the first stage, we train the entire architecture with frozen encoders and provide more appearance pixels to the generator. This allows the model to learn to disregard the correct facial expression and focus on generating faces that can fool the discriminators while implicitly encoding facial geometry in the appearance.

In the subsequent stages, we enforce the disentanglement of geometry and appearance by (1) enabling the encoder on sEMG occluded faces to update its weights and (2) gradually reducing the available appearance information. As a result, the model is forced to rely on the estimated facial geometry to restore correct faces over time.

Combining this multi-stage approach with the regularization terms introduced in the main paper ensures that the encoders correctly compute shape, expression, and position. This approach is crucial for achieving convergence, as it would otherwise require significantly more training effort.

A.6. EMG2Exp And Exp2EMG Architecture

We utilize simple multi-layer perceptrons (MLPs) to learn the non-linear relationship between the input data and the desired output for both our EMG2Exp (Synthesis) and Exp2EMG (Analysis) networks. These MLPscapture the complex relationships between the electromyography (EMG) signals and the corresponding facial expressions and vice versa. By employing MLPs, we effectively model the non-linear interactions between the input and output data.

As previously discussed, these models are trained in the second phase of EIFER, as the first phase requires a batch size of one. Therefore, we could not guarantee convergence of the training. By training them separately in the second phase, we can ensure that they learn the complex relationships between the input and output data effectively.

We provide a detailed illustration of both the *EMG2Exp* and the *Exp2EMG* models in Figure 4. In terms of architecture, we employ a simple yet effective design, utilizing `ReLU` activations [34] for all intermediate layers. This choice of activation function allows the models to learn non-linear relationships between the input and output data.

The final layer of each model is designed to accommodate the specific requirements of the output data. For the *EMG2Exp* model, we use a `Tanh` activation function, which allows the model to produce output values in the range of -1 to 1 (the typical ranges for the 3DMM expression space), suitable for representing facial expressions. In contrast, the *Exp2EMG* model uses a `ReLU` activation function in the final layer, as the sEMG signals are non-negative and require a non-negative output range.

During our experiments, we explored various expression encoder models, including DECA [14], EMOCAv2 [7], FOCUS [26], and Deep3DFace [42]. To accommodate the unique characteristics of each model, we adapted the input and output dimensions of our architecture accordingly, taking into account each model's specific expression parameter dimensions. This allowed us to effectively integrate these different expression encoder models into our framework and evaluate their performance in our experiments. Therefore, we compare the expression independently of the model architecture, gaining more insights into their underlying 3DMM and encoder capabilities instead.

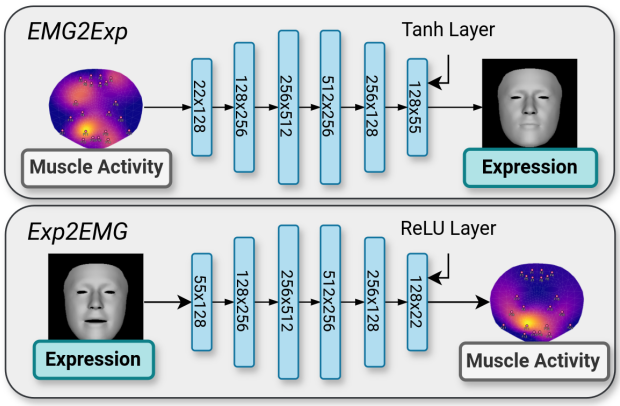


Figure 4. ***EMG2Exp* and *Exp2EMG* Architectures:** The simple MLP architecture learns the non-linear mapping between facial expression and muscle activity. Thus, the models learn the correspondence between these two domains.

B. Dataset - Mimics And Muscles

We created a custom dataset that simultaneously captures facial mimicry and muscle activity, bridging the gap between these two aspects. To our knowledge, this is the first dataset of its kind.

This dataset provides new insights into the complex dynamics between facial expressions and muscle activity. We provide a detailed description of the recording setup, recording scheme, surface electromyography schemes, data processing, and general data statistics to facilitate a deeper understanding of the dataset.

B.1. Recording Setup

In our experimental setup, a set of participants was instructed by an instruction video [48] to perform different facial movements. Each movement was repeated three times; thus, we can compare the repetitions against each other. Each movement task varied in time, ranging from 10 to 30 seconds. First, the following eleven facial movements were performed in that order:

1. Face-At-Rest
2. Forehead-Raise
3. Eye-Gentle
4. Eye-Tight
5. Smile-Closed
6. Smile-Open
7. Nose-Wrinkler
8. Cheeks Blow
9. Lip-Pucker
10. Snarl
11. Depress-Lip

Afterward, the participants had to mimic the six basic emotions [11] four times in total random order. The participants were shown faces to recreate. This ensured that no memory effect of previous repetitions could set in. Each expression was shown for three seconds, followed by a three-second interval for repetition. At 4.5 seconds, we assume the height peak during the expression.

We repeated the experiment twice with sEMG electrodes attached to measure muscle activity and once without electrodes as a reference. The duplicate sEMG measurement was conducted to ensure the reliability of the sEMG results. Additionally, we repeated the entire experiment two weeks later to account for potential changes in muscle activity and minimize inaccuracies that may arise from the participants' daily state. This allowed us to capture a more comprehensive and accurate representation of the participants' muscle activity over time.

Our participants were recorded with a frontal-facing Intel RealSence Depth Camera D415 (Intel Corporation, Santa Clara, California, U.S.) at 1280×720 resolution. Unfortunately, the obtained 3D information was unreliable and inaccurate in supporting the monocular 3D facial reconstruction but suitable enough for foreground and background separation.

We employ the same data collection setup as in [19, 36, 44]. To minimize skin impedance, all participants thoroughly cleaned their faces with non-refatting medical soap. The electromyography recording setup used reusable surface electrodes (Ag–Ag–Cl discs, diameter: 4 mm, DESS052606, GVB-geliMED, Bad Segeberg, Germany) to measure muscle activity. Reference electrodes (H93SG, Kendall, Germany) were bilaterally attached to the mastoid bone to provide a stable reference point. The muscle signals were amplified using sEMG amplifiers (ToEM16G, gain 100, frequency range 10–1,861 Hz, DeMeTec, Langgöns, Germany). Then they converted with an analog to digital converter (Tom, resolution: 5.96 nV/Bit, sampling rate: 4096/s, cutoff frequency: 2048 Hz, DeMeTec, Langgöns, Germany). The digitized data were then sampled using ATISArec (GJB Stentechnik, Ilmenau, Germany).

Our experimental setup allowed us to simultaneously record both the Fridlund [15] and Kuramoto [19, 25, 36] surface electromyography (sEMG) schemes. However, it is essential to note that the Kuramoto scheme provides regional information on muscle activity, whereas the Fridlund scheme offers more precise activation data. The electrode locations are illustrated in Figure 5, and our medical partners ensured accurate anatomic placement. A detailed description of the electrode channels is provided in Table 1, which reveals that some Fridlund electrodes overlap with Kuramoto electrodes in specific locations. For a more comprehensive understanding of the sEMG schemes and electrode placement, we refer the reader to previous studies [19, 36, 44].

While our primary focus is on the facial muscles responsible for expressions, we also recorded the activity of the *M. masseter*, a digestive muscle, and the *M. temporalis*. The facial muscles have been linked to specific facial movements through the Facial Action Coding System (FACS) [12], which is also included in Table 1. Notably, since we directly recorded facial expressions and muscle activity, we can bypass the intermediate Action Unit (AU) proxy variable in our approach. This unique aspect of our study offers benefits for improving and investigating the established FACS, providing new insights into the relationship between facial muscles and expressions.

B.2. Participant Cohort

We recruited 36 participants (19 ♀, 17 ♂, age range: 18-67 years) without a history of any neurological disease to obtain synchronous facial expression and muscle activity for this study. We specifically selected beardless male participants to ensure the accurate application of surface electromyography (sEMG) electrodes. Although our sample

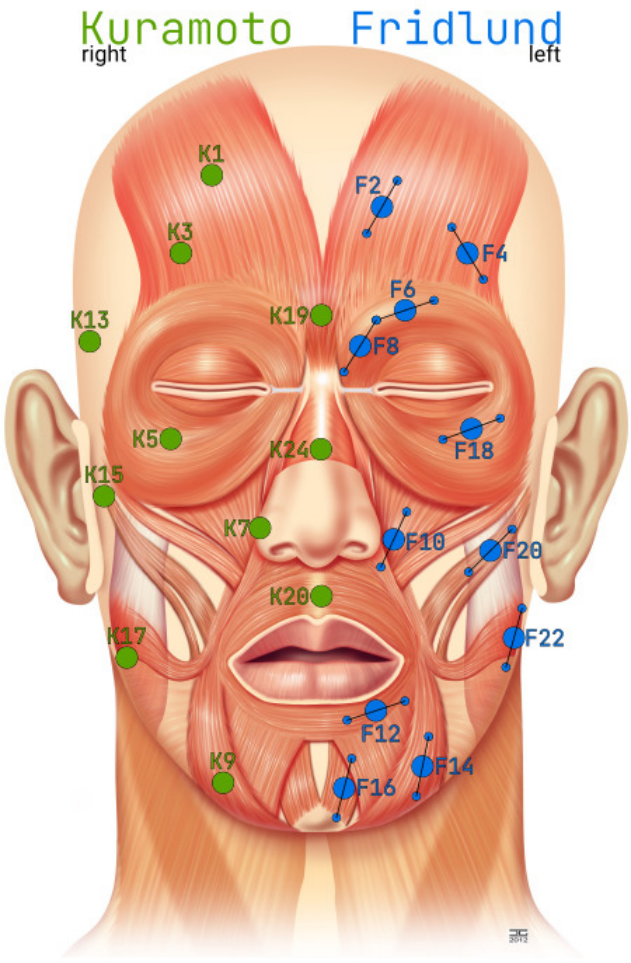


Figure 5. **sEMG Electrode Locations:** We highlight both surface electromyography schemes on their corresponding anatomical locations. Both Fridlund [15] (**F**, blue) and Kuramoto [25] (**K**, red) are attached to both face sides, but drawn here only on one for clarity. Please note that Fridlund is a bipolar scheme (denoted by the two smaller dots per electrode), and Kuramoto is monopolar by using **K24** as reference.

size is limited and may not represent an entire population, we aimed to achieve a balanced distribution of male and female participants across various age ranges. While the generalizability of our findings to a broader population remains uncertain, we expect the results to be consistent within this cohort, providing a reliable basis for further investigation.

To account for potential occlusions caused by the surface electromyography (sEMG) electrodes on key facial features, we recorded the same participants without electrode occlusion. This additional recording protocol allowed us to establish a reference dataset, which serves as a baseline for evaluating the accuracy of shape and expression reconstruction. By comparing the reconstructed results with the unoccluded recordings, we can assess the effectiveness

Fridlund	Kuramoto	Muscle	Action Unit	Movement
F1, F2	K1, K2	medialer frontalis	AU1	inner brow raiser
F3, F4	K3, K4	lateraler frontalis	AU2	outer brow raiser
F5, F6, F7, F8	K19	glabellae depressor supercilii corrugator supercilii	AU4	brow lowerer
F17, F18	K5, K6	orbicularis oculi	AU6	cheek raiser
F9, F10	K7, K8	levator labii superioris	AU9	nose wrinkler
F9, F10	K7, K8	levator labii superioris	AU10	upper lip raiser
F19, F20	-	zygomaticus minor	AU11	nasolabial deepener
F19, F20	(K15, K16)	zygomaticus major	AU12	lip corner puller
F13, F14	-	depressor anguli oris	AU15	lip corner depressor
F15, F16	K9, K10	mentalis	AU17	chin raiser
F11, F12	(K20)	philtrum, orbicularis oris	AU22	lip funneler
F11, F12	(K20)	orbicularis oris	AU23	lip tightener
F11, F12	(K20)	philtrum, orbicularis oris	AU24	lip pressor
F21, F22	K17, K18	masseter	AU26	jaw drop
F11, F12	-	philtrum, orbicularis oris	AU28	lip suck
-	K13, K14	temporalis	-	-

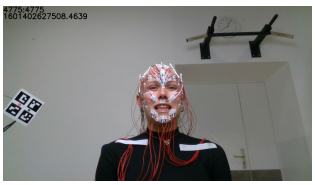
Table 1. **Electrode Channels and Muscles:** With our two sEMG electrode schemes, we capture the majority of facial muscles. We also included the according action units [12], providing insights into further research in the future. Please note channel names surrounded by brackets are just roughly attributable to the muscles, and **K11** and **K12** do not exist in the Kuramoto scheme [19, 25, 36, 44].

of our approach in capturing the nuances of facial expressions despite electrode placement.

B.3. Video Preprocessing

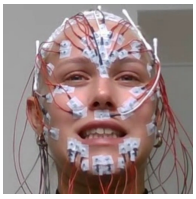
We provide a visualization of the original recording captured by the Intel RealSense camera, showcasing both RGB and depth data, in Figure 6 for a representative participant. To focus the model's attention on the most relevant facial regions, we employed the BlazeFace model [1] to compute the facial bounding box. However, not every frame yielded a valid bounding box, likely due to minor face orientation or unaccountable lighting changes. To address this, we interpolated missing bounding boxes using the position from the previous frame, assuming minimal participant movement due to the attached electrodes hindering a lot of movement. Additionally, Aruco markers placed on the left side of the frame facilitated synchronization across different data streams.

Following the extraction of bounding boxes, we leveraged rough depth information to segment the face from the background, thereby mitigating potential influences from external factors, such as people in the background. Subsequently, we applied a matting estimation technique using MODNet [22] to refine the segmentation results. Please note that the cables around the shoulder and neck area still make this segmentation challenging and might introduce artifacts. The outcome of this process is illustrated in Figure 6, also with artifacts above the right shoulder area. In conjunction with the recorded muscle activity data, these preprocessed frames were then utilized to train the EIFER model.

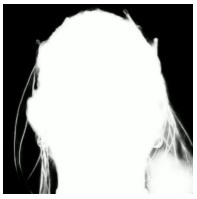


(a) Raw Input Video of Intel RealSense

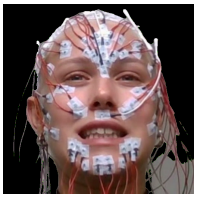
(b) Visualize Depth Map of the Intel RealSense



(c) Estimated Face Bouding Box via BlazeFace [1]



(d) Foreground masking via depth map and MODNet [22]



(e) Cropped and extracted face for the training

Figure 6. **Video Preprocessing:** We illustrate the preprocessing steps of the recorded facial videos for EIFER training, where we remove the background to facilitate facial expression extraction by the encoder models. Please note that you can see the Aruco markers on the left side of the raw input frame, which is used for the synchronization.

B.4. Electromyography Signal Preprocessing

We adhere to established protocols for processing the recorded electromyography (EMG) signals, as described in previous studies [16, 19, 36, 44, 50]. Specifically, we focus on the Fridlund scheme [15] due to its direct association with the corresponding muscles. As illustrated in Figure 7, our processing pipeline is uniformly applied to all sEMG recordings, including those for the *M. depressor anguli oris* (F19, F20) during the functional movement of *smiling*. The resulting signal exhibits the three repetitions of the movement. Notably, we refrain from normalizing the data during this preprocessing step, intentionally delaying normalization until the training phase to preserve participant-specific characteristics and avoid loss of information.

Unlike most research that typically operates on high-resolution sEMG (HR-sEMG) signals at 4096 Hz, we need to synchronize our signal with the recorded video at 30 frames per second (FPS). We employ a Fast Fourier Transform (FFT)-based downsampling approach [47], carefully ensuring that the essential frequency features are preserved. As the example demonstrates, the downsampling operation effectively maintains the signal's overall shape and fine-grained nuances. By successfully recording muscle activity and facial expression, we can explore the relationship between these two modalities, enabling a more comprehensive understanding of the underlying mechanisms.

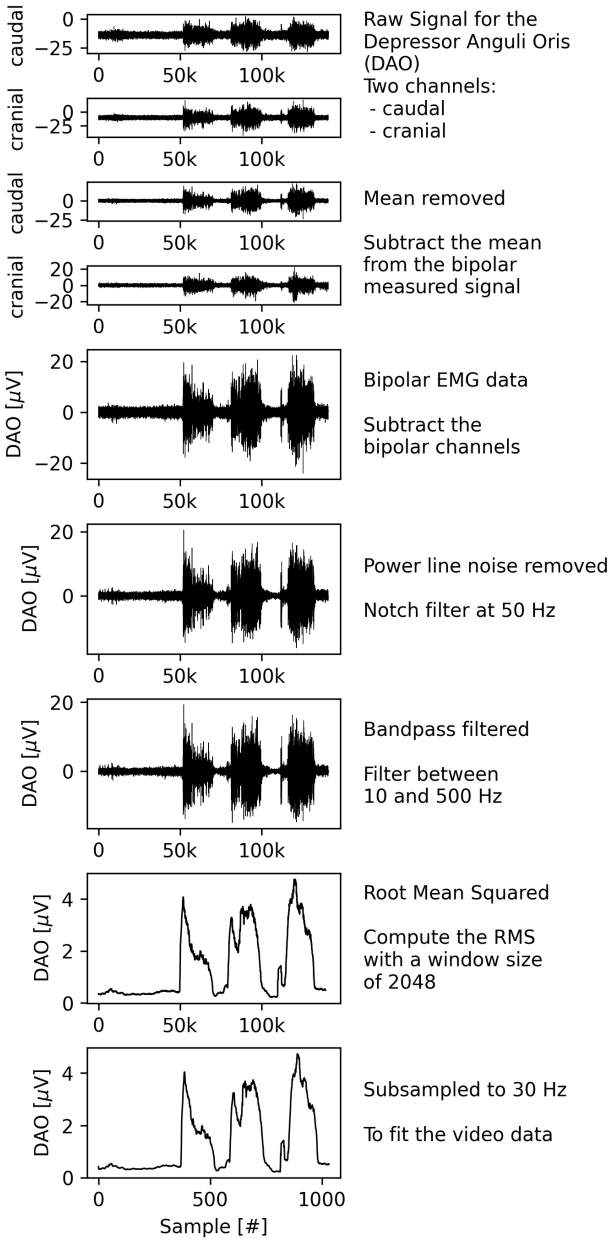


Figure 7. **Muscle Signal Preprocessing:** Illustration of the data processing pipeline for a single sEMG measurement of the *depressor anguli oris* muscle during the "Smiling" movement. Note the variability in the linear envelope of the measured muscle activity, even for repeated instances of the same movement.

B.5. Synchronization

We implemented an automated triggering system that simultaneously initiated both data streams to ensure precise synchronization between the video recording and surface electromyography (sEMG) signals. Additionally, we incorporated visual and sEMG-based synchronization triggers,

which were repeated twice to guarantee accurate alignment; see Figure 6 for the Aruco markers. This dual-triggering approach allowed us to align the video sections with the corresponding sEMG signals confidently. However, despite this rigorous synchronization protocol, some recordings still exhibited low confidence levels, necessitating their exclusion from the dataset. To provide transparency and account for these variations, we report the number of suitable recording snippets employed during training and evaluation for each facial movement in Table 2. This information promotes a more nuanced understanding of the dataset's composition and the reliability of our results.

Facial Movement	Total Recordings	Usable	Failed
Face-At-Rest	141	105	36
Forehead-Raise	141	106	35
Eye-Gentle	141	106	35
Eye-Tight	141	106	35
Cheeks-Blow	141	106	35
Smile-Closed	141	105	36
Smile-Open	141	106	35
Nose-Wrinkler	141	107	34
Lip-Pucker	141	106	35
Snarl	141	106	35
Depress-Lip	141	105	36
angry	560	528	32
disgusted	560	528	32
fearful	560	528	32
happy	560	528	32
sad	560	528	32
surprised	560	528	32
$\sum$	4911	4332	579

Table 2. **Synchronization Results:** We show how many recordings (at 30 FPS) of the synchronized facial expression and muscle activity are available for training and evaluation. Please note that the occlusion-free reference recording can be used fully.

B.6. Limitations

Our dataset is subject to several limitations that warrant consideration. Firstly, the facial expressions mimicked by participants may not accurately reflect natural, evoked expressions, as noted in previous studies [7, 11, 35]. However, this limitation does not compromise our ability to predict muscle activity from expressions and vice versa, as the measured facial muscle activity and recorded facial expressions still exhibit a strong correlation. This alignment ensures we can investigate the relationship between muscle activity and facial expressions despite the potential differences between mimicked and natural expressions.

Secondly, the surface electromyography (sEMG) electrodes used in our study introduce significant occlusion, which poses a challenge for feature extraction as illustrated in Figure 8. Existing methods are not trained on such data, and we cannot determine the potential bias these methods may introduce into our model [4, 5]. This highlights the need for more robust facial feature extraction to handle occlusions and ensure accurate predictions effectively.

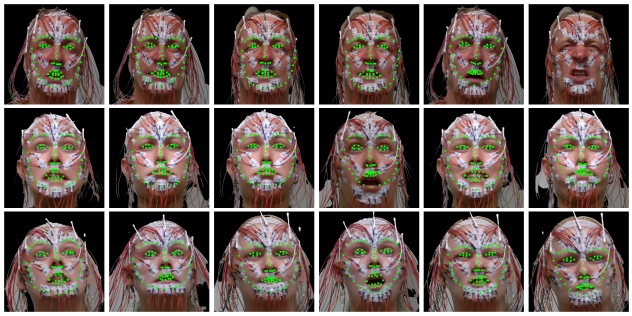


Figure 8. **Examples of Occluded Facial Expressions with Predicted Landmarks:** We present several examples of different facial expressions from three study participants. In addition to the original images, we overlay the predicted landmarks, as obtained using the methods described in [1, 29]. Notably, the predicted landmarks, if at all, exhibit inaccuracies, particularly in the outer regions of the face. This is of concern, as the outer part of the face is used to compute the convex hull for the masking function in SMIRK [40], and the strong offset observed in this region may impact the accuracy of the masking process.

Further, our recording setup is limited in capturing characteristic comprehensive muscle activity measurements. Specifically, certain muscle activities, such as the voluntary evoked eyelid closure, are controlled by the palpebral part of the *M. orbicularis oculi*, are not accounted for in the Fridlund scheme [15]. This omission is because the Fridlund scheme focuses on a specific set of facial muscles and the palpebral part of the *M. orbicularis oculi* is not included in this set. Additional measurements or specialized electrodes would be required to capture this activity, as discussed in previous studies [43]. This limitation highlights the need for more research concerning recording setups that can capture a broader range of muscle activities, enabling a more complete understanding of the complex relationships between facial muscles and expressions [16].

Our dataset has limitations, including its size and focus on functional movements, which may restrict the generalizability of our findings and the model's effectiveness in handling complex movements or subtle variations in facial expressions. Additionally, the impact of extreme facial expressions or diseases like facial palsy on our approach is unclear and warrants further investigation.

Our dataset was recorded within a medical study in Germany, subject to strict data privacy regulations. As a result, we are limited in the number of faces we can display

and the participants who agreed to share their data for new research databases. However, we will publish our trained models, *EMG2Exp* and *Exp2EMG*, which do not contain person-identifiable information, ensuring compliance with data protection regulations [10].

C. Experimental Setup

We compare EIFER to several state-of-the-art monocular 3D face reconstruction methods. Our comparison includes three models that employ the FLAME 3DMM [28]: DECA [14], EMOCAv2 [7], and SMIRK [40]. We also evaluate two models that use the BaselFaceModel [17, 38]: Deep3DFace [42] and FOCUS [26]. Additionally, we compare MC-CycleGAN [2, 3], which does not rely on a face model and implicitly learns the reconstruction. All models are available as PyTorch [37] implementations.

However, none of these methods were trained or tested on faces with sEMG electrodes attached. Moreover, our 36 participants were not part of their training data, making their faces completely unseen.

To ensure a fair comparison, we fine-tune all models on a common subset of occlusion-free reference recordings (10% of available frames). This approach has two benefits: First, it adapts the models to our data without occlusion, eliminating the need to account for their in-the-wild performance. Therefore, we assume they will perform best. Second, when applying the models to the sEMG-occluded faces of the same individuals, any behavior change can be attributed to the electrodes. This allows us to assess the models' invariance to this type of occlusion.

In contrast, EIFER trains on the same subset of occlusion-free faces (and uses the occluded faces, also 10% of available frames) as a reference to guide the reconstruction via adversarial challenge. As a result, all models have seen the same occlusion-free faces, making the comparison on the remaining 90% of frames fair.

We report the training hyperparameters for the first phase of EIFER, which focuses on expression reconstruction under sEMG occlusion.

We employ two AdamW [30] optimizers to train the encoder-generator pairs and discriminators independently. Both optimizers use a learning rate of $2 \cdot 10^{-4}$ and a weight decay of 10^{-3}. A cosine annealing learning rate scheduler adapts the learning rate during training. Again, we can only employ a batch size of one to facilitate the strength of instance normalization.

EIFER is trained for 20 epochs, divided into three stages: 10 epochs for the first, 5 for the second, and 5 for the last. We use 80% of the 10% available frames for training and 20% for validation. Note that the reported results in the main paper are on the 90% unseen frames.

During training, EIFER receives the triplet $(I^{\mathrm{N}}, I^{\mathrm{S}})$, where $I^{\mathrm{N}} \in \mathbb{R}^{224\times224\times3}$ is a color image of the occlusion-free face and $I^{\mathrm{S}} \in \mathbb{R}^{224\times224\times3}$ is a color image of the sEMG-occluded face. We apply random data augmentations to the frames, including random cropping, sharpening, and horizontal and vertical flipping.

During the second phase of EIFER, we train *EMG2Exp* and *Exp2EMG* using the following hyperparameters. We employ the Adam optimizer [24] with a learning rate of 10^{-3} and no additional learning rate scheduling or early-stopping. We use a batch size of 512 and train for 200 epochs. All results in the main paper are reported on a five-fold cross-validation.

Both models are trained on the tuple (A, φ), where $A \in \mathbb{R}^{22}$ represents the 22 measured muscle signals using the Fridlund sEMG scheme. We normalize the muscle signals A by the maximum measured muscle activity *for each participant.* This normalization accounts for individual intensity and muscle strength differences, allowing for a more comparable analysis across participants. Please note that this maximum value has been used to restore the reconstructed activity during the $Exp2EMG$ predictions. φ denotes the 3DMM expression space parameters. The dimension of φ varies across models:

- For EIFER and SMIRK [40], $\varphi \in \mathbb{R}^{55}$ (50 expressions, two eyelids, three jaw).
- For DECA [14] and EMOCAv2 [7], $\varphi \in \mathbb{R}^{53}$ (50 expression parameters and three jaw).
- For FOCUS, $\varphi \in \mathbb{R}^{100}$.
- For Deep3DFace, $\varphi \in \mathbb{R}^{64}$.

Please note that FLAME [28] models the jaw movement intentionally separate, and BFM [17, 38] models this implicitly via the expression space. This allows us to compare the expression space differences between FLAME [28] and BaselFaceModel [17, 38].

D. Visualizations And Videos

We provide additional visual examples for each main experiment, including videos to highlight our approach's dynamic aspects and highlight our methods' advantages.

D.1. Isolated Shape Visualization

In our experiments, we observed that the same individual was reconstructed with varying facial geometries. To investigate this, we analyzed the shape parameters of both FLAME [28] and BFM [17, 38] under neutral expressions, excluding camera or pose parameters. Notably, EMOCAv2 [7] employs the same encoder as DECA [14], resulting in identical shape parameter estimates. Our analysis revealed that all models, except EIFER, exhibited differences in geometry for the same individual. This discrepancy may explain why the expression parameters, compensating for the visual reconstruction, potentially affect the quality of muscle activation predictions in our later experiments.

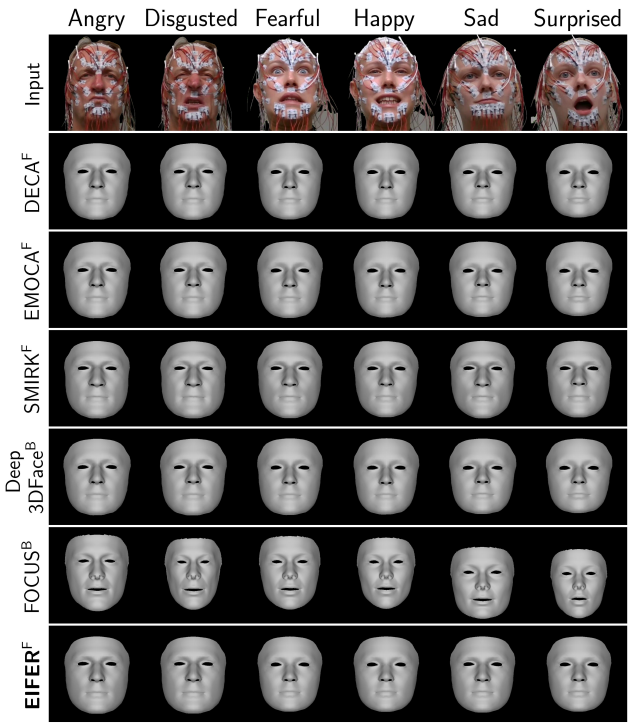


Figure 9. Isolated *shape* parameters of the facial reconstruction. Many models have slightly different shape geometries for the same individual, indicating that the encoder might use the expression space to substitute the reconstruction.

D.2. Reconstruction

We provide additional visual examples for facial geometry extraction and appearance reconstruction. We also demonstrate the reconstruction using only expression parameters on a neutral face to evaluate the encoder's disentanglement ability during sEMG occlusion. Examples include Face-At-Rest, Eye-Tight, Smile-Open, Snarl, and Nose-Wrinkler. These are the following figures:

- Face-At-Rest: 3D Geometry Figure 10
- Face-At-Rest: Isolated Expression Figure 11
- Face-At-Rest: Appearance Reconstruction Figure 12
- Eye-Tight: 3D Geometry Figure 13
- Eye-Tight: Isolated Expression Figure 14
- Eye-Tight: Appearance Reconstruction Figure 15
- Smile-Open: 3D Geometry Figure 16
- Smile-Open: Isolated Expression Figure 17
- Smile-Open: Appearance Reconstruction Figure 18
- Snarl: 3D Geometry Figure 19
- Snarl: Isolated Expression Figure 20
- Snarl: Appearance Reconstruction Figure 21
- Nose-Wrinkler: 3D Geometry Figure 22
- Nose-Wrinkler: Isolated Expression Figure 23
- Nose-Wrinkler: Appearance Reconstruction Figure 24

D.3. EMG2Exp

We provide additional visual examples of synthesized facial expressions based on muscle activity for all methods, including the six base emotions and eleven functional movements for more participants, as shown in Figure 25. We also compare the results using MC-CycleGAN [2, 3] restored recordings for a fair comparison.

Our method directly generates highly realistic faces from occluded faces, whereas other methods require occlusion-free faces. This demonstrates the robustness of EIFER in handling sEMG occlusion. However, SMIRK [40] is the only method to reconstruct the *Depress-Lip* movement, demonstrating its ability to encode rare and subtle facial expressions. In contrast, EIFER could not learn this movement, even under occlusion, highlighting a potential area for improvement.

We observe an interesting phenomenon where the model can synthesize the *Eye-Tight* movement but not the *Eye-Gentle* movement. This suggests that the model can pick up on different muscular patterns depending on the strength of the same movement. However, it remains unclear whether the differences between voluntary and enforced movements exhibit similar patterns. Notably, EIFER is the only method that can restore the *Lip-Pucker* movement.

We also observe that the jaw movement is challenging to learn, as the *M. massester* muscle is only slightly active during jaw opening. Although this task is easy to solve visually, the muscle activity appears insufficient. Furthermore, we find that the performance of the two 3DMMs (FLAME [28] and BaselFaceModel [17, 38]) depends on the encoder model. This suggests that a well-trained encoder model is more important than the capabilities of the 3DMM expression space.

This finding highlights the importance of disentanglement of shape and expression in 3DMMs, as well as the significance of the encoder model [9, 10]. Although this task remains ill-posed, our results have implications for new research directions in medicine and psychology.

D.4. Exp2EMG

We provide additional examples of EIFER's muscle activity prediction beyond the single active and inactive muscle visualized in the main paper for the *happy* expression. These can be found in Figure 26, Figure 27, and Figure 28. Certain muscles are typically active during specific facial expressions, while others remain inactive. However, we also notice decreased activity in some muscles, accompanied by activation in others. This phenomenon, which is not well-studied [19], suggests that facial muscles may be more interconnected than currently assumed [11, 12], warranting further investigation.

EIFER can accurately predict the muscle activity envelope without requiring additional personal information

However, we refine the prediction by multiplying it by the participant's maximum observed activity (in μV), allowing us to estimate the relative activity and actual muscle strength. Even without this refinement, EIFER remains a powerful tool for predicting muscle activity.

We observe that EIFER accurately fits the shape of the original signal in all reconstructions but occasionally struggles to estimate the signal amplitude correctly. We attribute this to the per-participant normalization during training, which may cause the model to underestimate the general signal amplitude if participants require varying levels of muscle activity to evoke changes in facial mimicry.

Several potential reasons for this phenomenon deserve further exploration:

1. Are there differences in voluntary and evoked expression patterns?
2. Do participants exhibit unique muscle activity patterns for certain expressions due to pathological conditions?
3. Are there learning effects between sessions, such as changes in reaction time, execution speed, or intensity?

To drive progress in understanding and addressing these open questions, we are releasing our models *EMG2Exp* and *Exp2EMG* to the research community, inviting collaboration and exploration to uncover the underlying causes of these phenomena and push the boundaries of facial expression analysis.

Figure 10. Facial Geometry Reconstruction during **Face-At-Rest**

Figure 11. Isolated Facial Expression Reconstruction during **Face-At-Rest**

Figure 12. Facial Geometry Reconstruction during **Face-At-Rest**

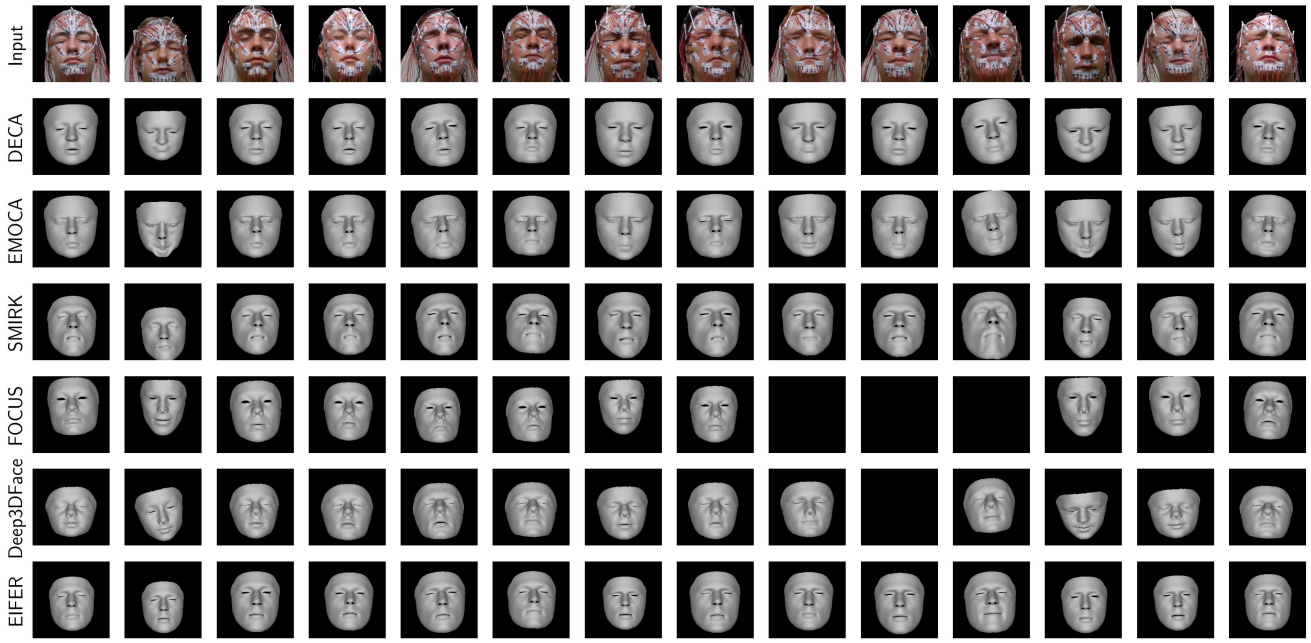


Figure 13. Facial Geometry Reconstruction during **Eye-Tight**

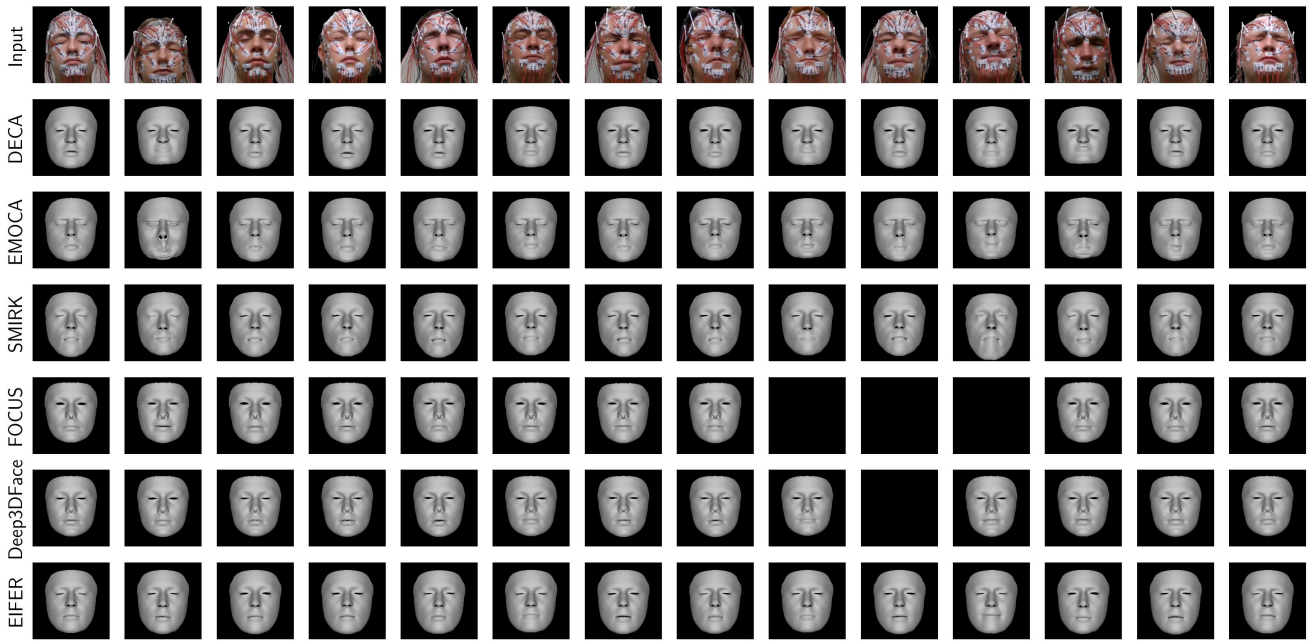


Figure 14. Isolated Facial Expression Reconstruction during **Eye-Tight**

Figure 15. Facial Geometry Reconstruction during **Eye-Tight**

Figure 16. Facial Geometry Reconstruction during **Smile-Open**

Figure 17. Isolated Facial Expression Reconstruction during **Smile-Open**

Figure 18. Facial Geometrie Reconstruction during **Smile-Open**

Figure 19. Facial Geometry Reconstruction during **Snarl**

Figure 20. Isolated Facial Expression Reconstruction during **Snarl**

Figure 21. Facial Geometry Reconstruction during **Snarl**

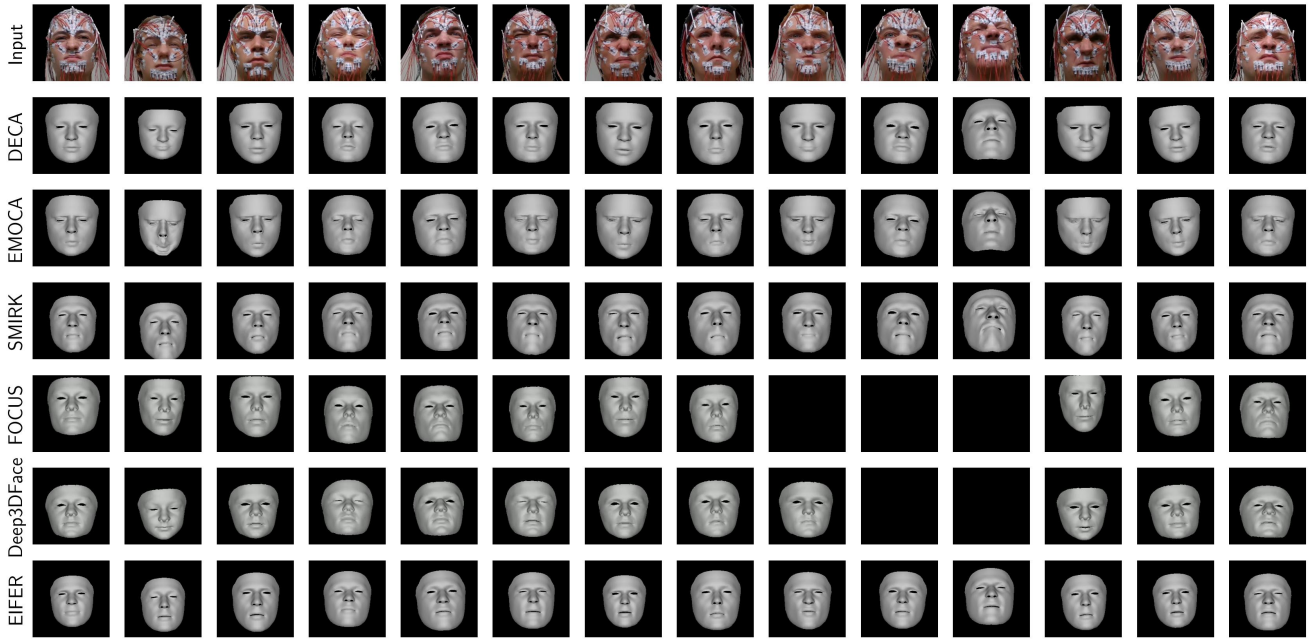


Figure 22. Facial Geometry Reconstruction during **Nose-Wrinkler**

Figure 23. Isolated Facial Expression Reconstruction during **Nose-Wrinkler**

Figure 24. Facial Geometry Reconstruction during **Nose-Wrinkler**

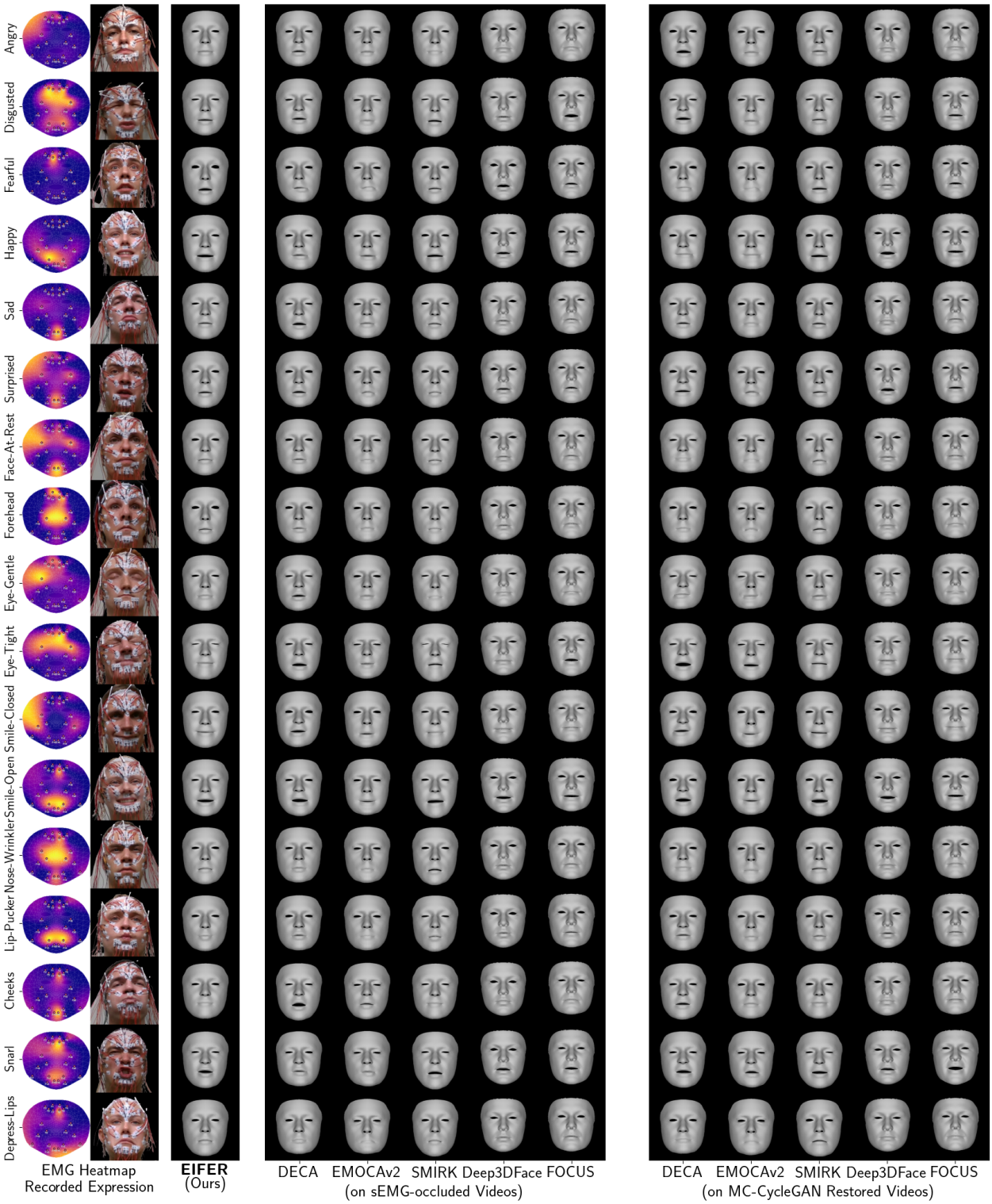


Figure 25. **Physiological-based Expression Synthesis via Muscle Activity:** We demonstrate synthesized facial expressions from recorded muscle activity. State-of-the-art methods, such as SMIRK and FOCUS, struggle to reconstruct expressions under sEMG occlusion. We see improved results on the MC-CycleGAN [2, 3] restored faces, but only SMIRK performs well across all emotions. In contrast, our method, EIFER, achieves comparable synthesis quality directly from occluded images without needing electrode removal.

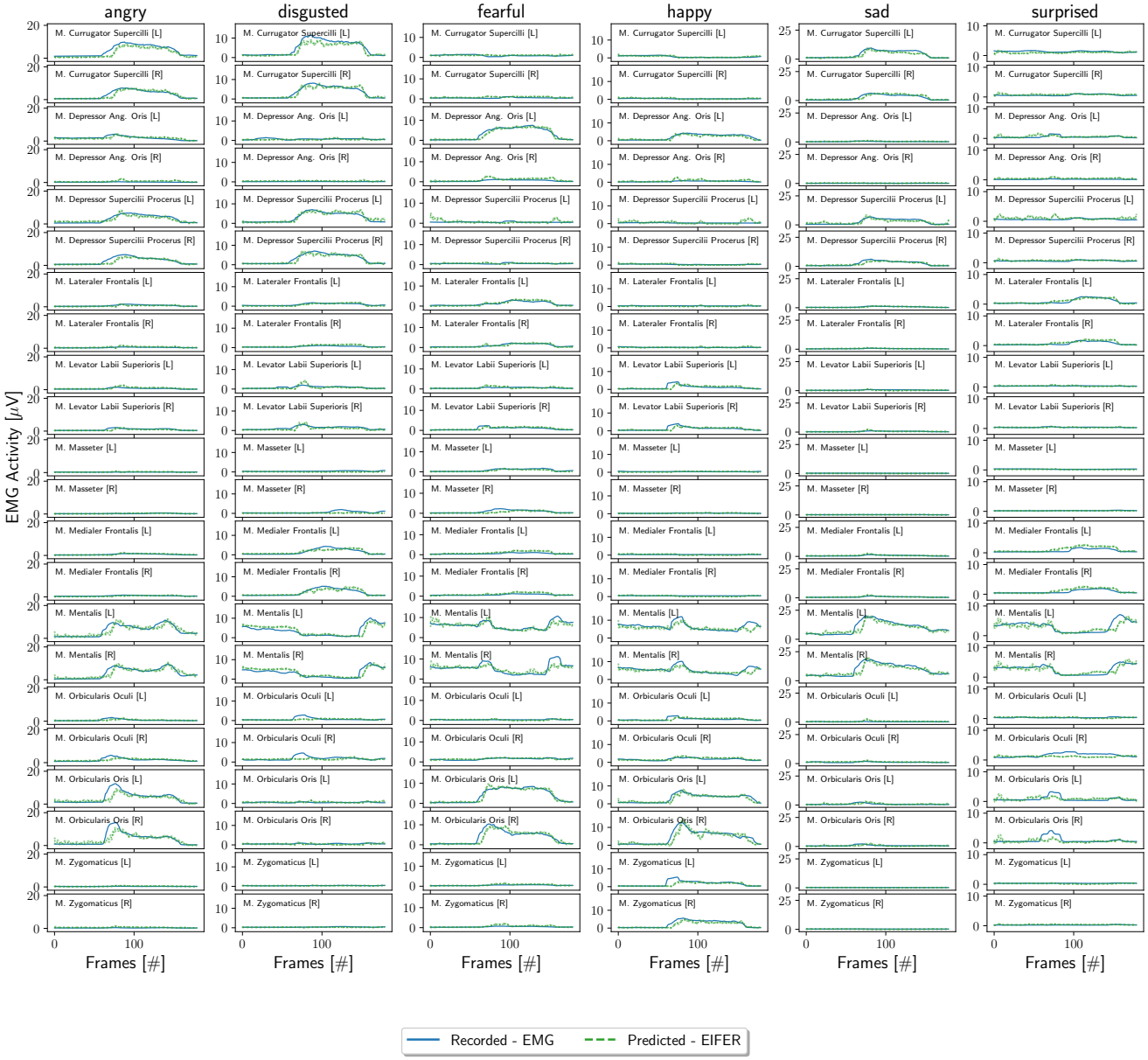


Figure 26. **Muscle Activity via Expression Parameters** We demonstrate the reconstruction of muscle activity from expression parameters, achieving fair results with minor amplitude signal issues. We visualize this capability for the six base emotions [11].

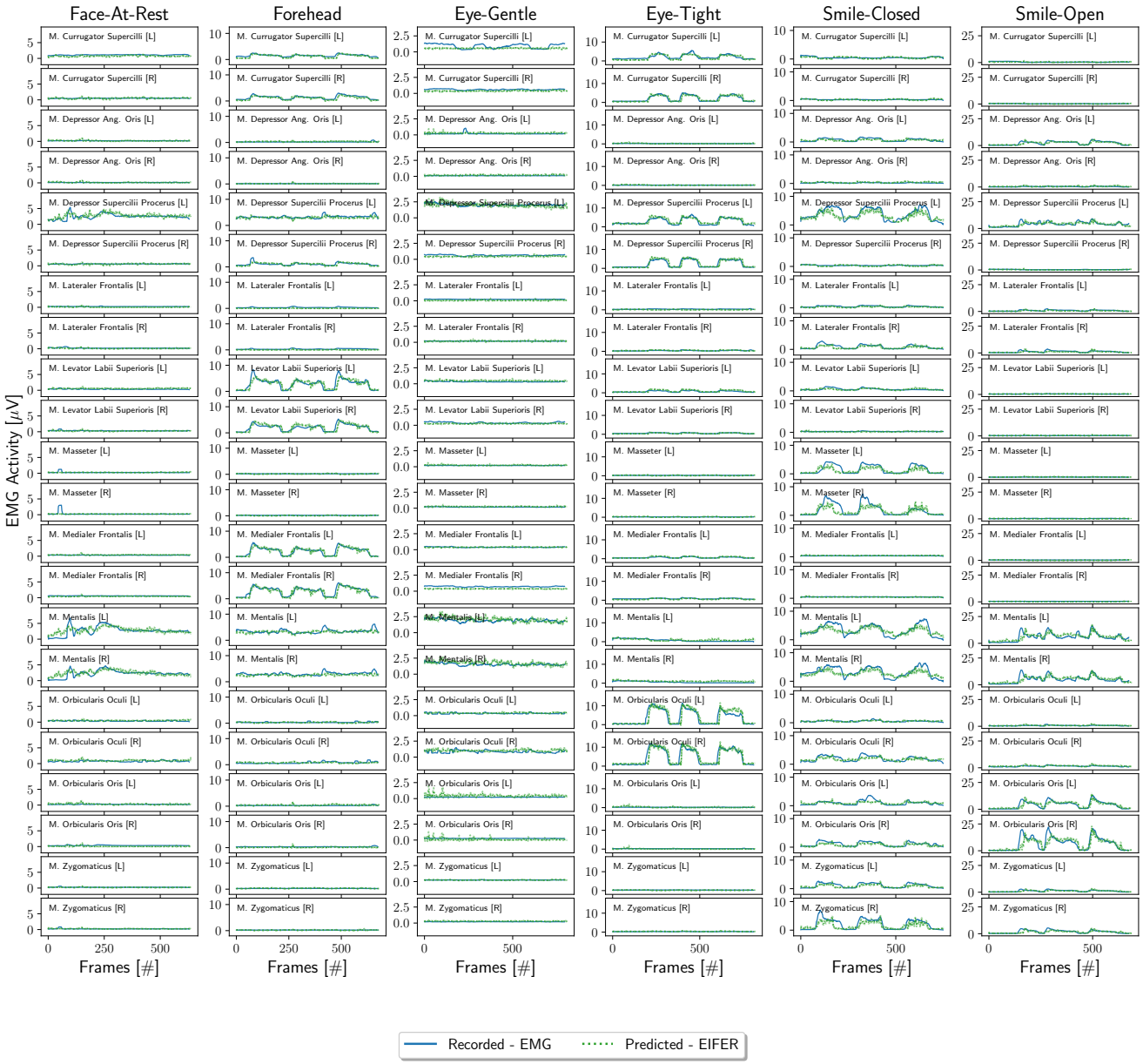


Figure 27. **Muscle Activity via Expression Parameters** We demonstrate the reconstruction of muscle activity from expression parameters, achieving fair results with minor amplitude signal issues. We visualize this capability for the six different functional movements [48].

Figure 28. **Muscle Activity via Expression Parameters** We demonstrate the reconstruction of muscle activity from expression parameters, achieving fair results with minor amplitude signal issues. We visualize this capability for the remaining five functional movements [48].

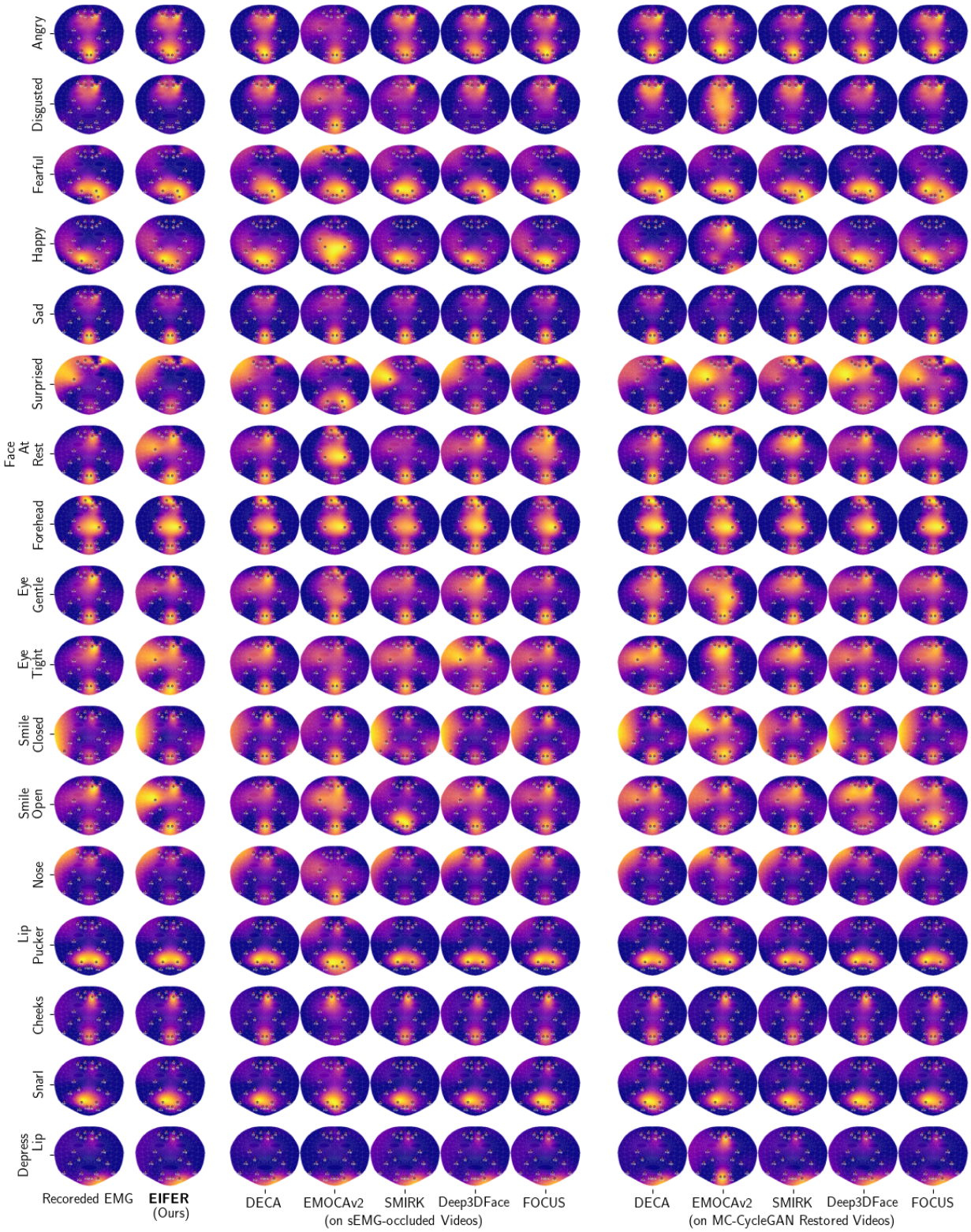


Figure 29. **Topological EMG Heatmaps:** We compare the topological heat maps during the peak muscle measurement of muscle activity for each movement. Further, we display the predicted muscle activity based on each method. SI units are committed for clarity but can be taken from the other muscle activity prediction figures (Figure 26, Figure 27, Figure 28).

E. Ablation Studies

In addition to the results presented in the main paper, we conduct further ablation studies to explore alternative applications of EIFER. We investigate various downstream tasks to assess the model's versatility and potential uses beyond its original purpose.

E.1. Convolutional Based Expression Classification

We leverage the six base emotions [11] mimicked by our participants as reference annotations, serving as ground truth labels for image-based classification. This way, we evaluate whether the appearance reconstruction accurately resembles the target facial expression, providing an additional objective criterion for assessing appearance quality.

We employ several convolution-based Facial Expression Recognition (FER) classifiers: Poster++ [32], ResidualMaskingNet [31], EmoNext [13], and SegmentationVGG [46]. All required preprocessing steps are strongly followed, as outlined in the paper and corresponding code repositories. While this model selection is not exhaustive, it gives a broad overview of existing classifiers trained on public datasets. However, since we cannot directly assess the accuracy of the mimicked expression, we establish two baselines: (1) an upper baseline using occlusion-free reference recordings, and (2) a lower threshold using sEMG-occluded recordings. Any model should perform better than the lower baseline.

We present the results in Table 3 to Table 6. Notably, none of the reconstruction methods achieve the original upper limit on the occlusion-free videos. This discrepancy may be attributed to several factors: First, the methods may struggle with frames that differ from the training database's image quality and recording style, simulating a distribution shift or in-the-wild application scenario. Second, the appearance reconstruction may introduce biases invisible to the human eye but affect the models' performance [4, 5]. Lastly, the reconstruction may not retain the essential facial features that the models rely on, indicating potential information loss during the reconstruction. While the underlying cause is beyond the scope of this study, it is an essential area of research that can help uncover the black-box nature of FER classification models.

E.2. Landmarks under Occlusions

We demonstrate in Figure 8 that existing landmarking models struggle to predict landmarks accurately under sEMG occlusion. However, EIFER, trained without landmark information, still aligns well with the facial geometry. We leverage this alignment to predict landmarks, as defined on the FLAME model [14, 28, 40]. Although we lack groundtruth annotations for the landmarks, visual inspection reveals that EIFER's predictions outperform those of existing models [1, 29], as shown in Figure 30. While

	Angry	Disgusted	Fearful	Happy	Sad	Surprised	Average
Upper Limit (N)	65.97	82.29	53.12	94.08	75.00	70.49	73.49
Lower Limit (S)	14.02	15.91	46.02	54.92	84.47	60.04	45.90
DECA	7.01	0.00	0.00	0.00	0.00	21.97	4.83
EMOCAv2	61.45	3.39	79.89	2.82	8.94	6.15	27.11
SMIRK	31.44	4.92	22.73	56.06	81.06	53.03	41.54
Deep3DFace	0.25	1.50	5.46	32.17	16.42	89.58	24.23
FOCUS	0.26	0.00	0.00	1.56	0.00	92.23	15.67
MCGAN	47.97	75.61	33.94	75.61	73.17	58.94	60.87
EIFER	48.67	61.55	27.84	71.78	67.23	56.44	55.59

Table 3. **Emotion Classification Accuracy for Poster++[32]:** We report the FER image-based classification results for the appearance reconstructions.

	Angry	Disgusted	Fearful	Happy	Sad	Surprised	Average
Upper Limit (N)	56.60	72.57	36.81	84.67	8.33	57.29	52.71
Lower Limit (S)	13.64	0.00	2.65	25.19	0.00	81.82	20.55
DECA	0.19	0.00	82.01	0.57	0.00	0.00	13.79
EMOCAv2	2.23	7.91	4.47	3.95	1.12	1.12	3.47
SMIRK	16.10	4.73	5.68	7.77	0.57	68.37	17.20
Deep3DFace	4.96	15.54	3.97	60.60	2.24	68.98	26.05
FOCUS	2.07	1.31	40.57	18.44	0.26	47.15	18.30
MCGAN	45.12	64.43	21.14	52.85	2.64	39.63	37.64
EIFER	48.67	62.69	14.20	55.11	3.22	39.39	37.22

Table 4. **Emotion Classification Accuracy for ResidualMaskingNet[31]:** We report the FER image-based classification results for the appearance reconstructions.

	Angry	Disgusted	Fearful	Happy	Sad	Surprised	Average
Upper Limit (N)	48.96	0.00	14.24	97.21	21.88	69.44	41.95
Lower Limit (S)	19.32	0.00	30.68	61.93	31.82	34.28	29.67
DECA	0.00	0.00	0.00	0.19	11.74	5.11	2.84
EMOCAv2	60.89	0.00	16.20	6.21	51.40	48.04	30.46
SMIRK	70.27	0.00	17.42	60.61	13.83	9.28	28.57
Deep3DFace	1.24	0.00	3.97	65.34	4.23	66.75	23.59
FOCUS	0.00	0.00	0.00	20.52	0.26	35.49	9.38
MCGAN	21.54	0.00	5.89	90.24	8.94	43.70	28.39
EIFER	48.11	0.00	2.65	94.70	8.52	35.80	31.63

Table 5. **Emotion Classification Accuracy for EmoNextBase[13]:** We report the FER image-based classification results for the appearance reconstructions. Please note that the model has never predicted *disgust* for any image.

	Angry	Disgusted	Fearful	Happy	Sad	Surprised	Average
Upper Limit (N)	21.53	0.00	11.46	78.75	69.10	1.74	30.43
Lower Limit (S)	16.86	0.00	5.49	71.97	22.35	0.19	19.48
DECA	0.19	0.00	0.19	0.38	5.30	3.98	1.67
EMOCAv2	63.69	0.00	10.06	18.08	31.28	52.51	29.27
SMIRK	38.83	0.00	2.65	45.45	24.81	25.57	22.89
Deep3DFace	0.99	0.00	1.99	40.40	20.65	48.14	18.69
FOCUS	1.81	0.00	0.52	31.43	1.30	40.93	12.67
MCGAN	51.42	0.00	2.85	69.11	14.84	12.20	25.07
EIFER	61.74	0.00	1.70	72.54	10.04	15.72	26.96

Table 6. **Emotion Classification Accuracy for SegmentationVGG19[46]:** We report the FER image-based classification results for the appearance reconstructions. While SegmentationVGG19 performs well on the benchmark datasets, the application to our unseen data results in strong performance degradation.

EIFER shows improved alignment, there is still room for improvement.

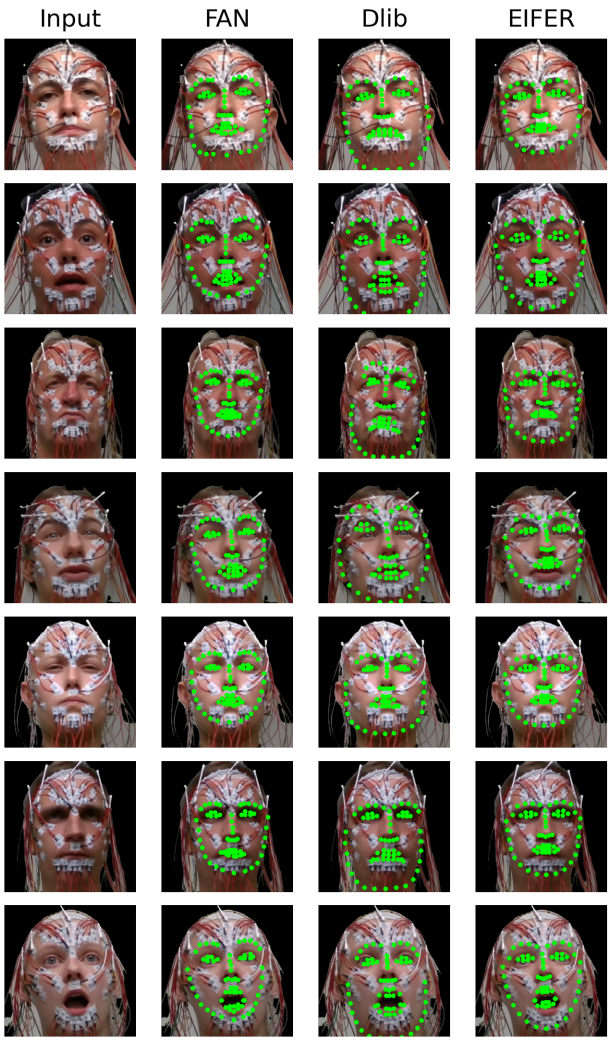


Figure 30. **Landmark Prediction under sEMG occlusion:** We see that EIFER can be used to predict the 2D facial landmarks under occlusion, whereas existing methods [23, 29] produce inaccurate predictions.

F. Extended Limitations Discussion

Facial Action Coding System: EIFER presents a novel data-driven approach for estimating muscle signals from facial expressions using electromyography. Comparing this paradigm to traditional methods, such as the Facial Action Coding System (FACS) [12], is an open research direction that we leave for future work. Existing FACS regression methods do work on our occluded images. Instead, we rely on MC-CycleGAN recordings [2, 3] to make a fair comparison. However, as shown in Section E.1, the appearance reconstructions of these models differ from occlusion-free reference recordings. Further research is necessary to ensure the suitability of our dataset for a comprehensive comparison study.

Generalization: Our results' generalizability is uncertain due to the limited sample size ($N = 36$). Additionally, our cohort is based in Germany, which may introduce cultural biases that could impact the results when applied to other populations. We tested a wide range of standardized facial expressions [11, 48], but participants did not perform them voluntarily. This may affect the generalizability of spontaneous facial expressions, which might exhibit different muscle activity patterns. However, our results still captured facial mimicry and muscle activity, suggesting that the learned correspondence remains valid. Our study only includes healthy participants without pre-existing neurological diseases affecting the facial nerve. Therefore, conditions like facial palsy or Parkinson's disease may impact the predictions. EIFER may not address facial asymmetry typical in facial palsy, as it may have learned a symmetry bias from our data [4, 5]. Furthermore, our models might not recover synkinetic effects (involuntary movements on the contralateral face side). To address this, we currently record patients with unilateral synkinetic chronic facial palsy to validate our approach for medical use cases.

Data Availability: Our dataset was recorded in Germany as part of a medical study, subject to strict data privacy regulations. Due to these regulations, we are restricted in sharing participant data and faces. However, we will release our trained models, *EMG2Exp* and *Exp2EMG*, which do not contain person-identifiable information, ensuring compliance with data protection regulations [10].

Disentanglement: Our approach relies on the disentanglement of shape and expression in 3D Morphable Models (3DMMs), specifically FLAME [28] and BaselFace-Model [17, 38], as well as the face encoder's ability to establish this correspondence [10, 49]. The behavior of other 3DMMs, such as FaceScapes [51, 52], ICT-FaceKIT [27], or FaceWarehouse [6], is unclear and requires further investigation. A necessary condition for exploring these models is the availability of well-pre-trained encoder models. Without these, the correspondence between facial expressions and muscle activity might not be learnable.

References - Supplementary

[Sup1] Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*, 2019.

[Sup2] Tim Büchner, Orlando Guntinas-Lichius, and Joachim Denzler. Improved obstructed facial feature reconstruction for emotion recognition with minimal change cyclegans. In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pages 262–274. Springer, 2023.

[Sup3] Tim Büchner, Sven Sickert, Gerd Fabian Volk, Christoph Anders, Orlando Guntinas-Lichius, and Joachim Denzler. Let's get the facs straight–reconstructing obstructed facial features. *arXiv preprint arXiv:2311.05221*, 2023.

[Sup4] Tim Büchner, Niklas Penzel, Orlando Guntinas-Lichius, and Joachim Denzler. Facing asymmetry–uncovering the causal link between facial symmetry and expression classifiers using synthetic interventions. *arXiv preprint arXiv:2409.15927*, 2024.

[Sup5] Tim Büchner, Niklas Penzel, Orlando Guntinas-Lichius, and Joachim Denzler. The power of properties: Uncovering the influential factors in emotion classification. *arXiv preprint arXiv:2404.07867*, 2024.

[Sup6] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. FaceWarehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[Sup7] Radek Danecek, Michael J Black, and Timo Bolkart. EMOCA: Emotion Driven Monocular Face Capture and Animation. *CVPR*, page 12, 2022. 8

[Sup8] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019.

[Sup9] Bernhard Egger, William A. P. Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, Christian Theobalt, Volker Blanz, and Thomas Vetter. 3D Morphable Face Models-Past, Present, and Future. *ACM Transactions on Graphics*, 39(5):157:1–157:38, 2020.

[Sup10] Bernhard Egger, Skylar Sutherland, Safa C Medin, and Joshua Tenenbaum. Identity-expression ambiguity in 3d morphable face models, 2021.

[Sup11] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[Sup12] Paul Ekman and W Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Palo Alto: Consulting Psychologists Press*, 1978.

[Sup13] Yassine El Boudouri and Amine Bohi. EmoNeXt: An Adapted ConvNeXt for Facial Emotion Recognition. In *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2023.

[Sup14] 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*, 40(4):1–13, 2021. 8

[Sup15] Alan J. Fridlund and John T. Cacioppo. Guidelines for human electromyographic research. *Psychophysiology*, 23 (5):567–589, 1986.

[Sup16] Paul F. Funk, Bara Levit, Chen Bar-Haim, Dvir Ben-Dov, Gerd Fabian Volk, Roland Grassme, Christoph Anders, Orlando Guntinas-Lichius, and Yael Hanein. Wireless high-resolution surface facial electromyography mask for discrimination of standardized facial expressions in healthy adults. *Scientific Reports*, 14(1):19317, 2024.

[Sup17] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schoenborn, and Thomas Vetter. Morphable face models - an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82, 2018. 8

[Sup18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Sup19] Orlando Guntinas-Lichius, Vanessa Trentzsch, Nadiya Mueller, Martin Heinrich, Anna-Maria Kuttenreich, Christian Dobel, Gerd Fabian Volk, Roland Graßme, and Christoph Anders. High-resolution surface electromyographic activities of facial muscles during the six basic emotional expressions in healthy adults: A prospective observational study. *Scientific Reports*, 13(1):19214, 2023.

[Sup20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Sup21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[Sup22] Zhanghan Ke, Jiayu Sun, Kaican Li, Qiong Yan, and Rynson W.H. Lau. Modnet: Real-time trimap-free portrait matting via objective decomposition. In *AAAI*, 2022.

[Sup23] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[Sup24] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2017.

[Sup25] Eriko Kuramoto, Saori Yoshinaga, Hiroyuki Nakao, Seiji Nemoto, and Yasushi Ishida. Characteristics of facial muscle activity during voluntary facial expressions: Imaging analysis of facial expressions based on myogenic potential data. *Neuropsychopharmacology Reports*, 39(3): 183–193, 2019.

[Sup26] Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. Robust model-

based face reconstruction through weakly-supervised outlier segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 372–381, 2023.

[Sup27] Ruilong Li, Karl Bladin, Yajie Zhao, Chinmay Chinara, Owen Ingraham, Pengda Xiang, Xinglei Ren, Pratusha Prasad, Bipin Kishore, Jun Xing, et al. Learning formation of physically-based face attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3410–3419, 2020.

[Sup28] 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*, 36(6):1–17, 2017. 8

[Sup29] Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. Dense face alignment, 2017.

[Sup30] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, 2019.

[Sup31] Pham Luan, Vu Huynh, and Tran Tuan Anh. Facial expression recognition using residual masking network. In *IEEE 25th International Conference on Pattern Recognition*, pages 4513–4519, 2020.

[Sup32] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, Aibin Huang, and Yigang Wang. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, page 110951, 2024.

[Sup33] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks, 2017.

[Sup34] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[Sup35] 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, 2019.

[Sup36] Nadiya Mueller, Vanessa Trentzsch, Roland Grassme, Orlando Guntinas-Lichius, Gerd Fabian Volk, and Christoph Anders. High-resolution surface electromyographic activities of facial muscles during mimic movements in healthy adults: A prospective observational study. *Frontiers in Human Neuroscience*, 16:1029415, 2022.

[Sup37] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[Sup38] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 296–301, Genova, Italy, 2009. IEEE. 8

[Sup39] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020.

[Sup40] George Retsinas, Panagiotis P Filntisis, Radek Danecek, Victoria F Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 3d facial expressions through analysis-by-neural-synthesis, 2024.

[Sup41] 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 18*, pages 234–241. Springer, 2015.

[Sup42] Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. Self-supervised monocular 3d face reconstruction by occlusion-aware multi-view geometry consistency. In *European Conference on Computer Vision*, pages 53–70. Springer, 2020.

[Sup43] Katharina Steiner, Marius Arnz, Gerd Fabian Volk, and Orlando Guntinas-Lichius. Electro-stimulation system with artificial-intelligence-based auricular-triggered algorithm to support facial movements in peripheral facial palsy: A simulation pilot study. *Diagnostics*, 14(19):2158, 2024.

[Sup44] Vanessa Trentzsch, Nadiya Mueller, Martin Heinrich, Anna-Maria Kuttenreich, Orlando Guntinas-Lichius, Gerd Fabian Volk, and Christoph Anders. Test-retest reliability of high-resolution surface electromyographic activities of facial muscles during facial expressions in healthy adults: A prospective observational study. *Frontiers in Human Neuroscience*, 17:1126336, 2023.

[Sup45] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *ArXiv*, abs/1607.08022, 2016.

[Sup46] S. Vignesh, M. Savithadevi, M. Sridevi, and Rajeswari Sridhar. A novel facial emotion recognition model using segmentation VGG-19 architecture. *International Journal of Information Technology*, 15(4):1777–1787, 2023.

[Sup47] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[Sup48] Gerd Fabian Volk, Rebecca Anna Schaede, Jovanna Thielker, Luise Modersohn, Oliver Mothes, Charles C.

Nduka, Jodi Maron Barth, Joachim Denzler, and Orlando Guntinas-Lichius. Reliability of grading of facial palsy using a video tutorial with synchronous video recording. *The Laryngoscope*, 129(10):2274–2279, 2019.

[Sup49] Maximilian Weiherer, Finn Klein, and Bernhard Egger. Approximating Intersections and Differences Between Linear Statistical Shape Models Using Markov Chain Monte Carlo. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6352–6361, Waikoloa, HI, USA, 2024. IEEE.

[Sup50] Peng Xia, Jie Hu, and Yinghong Peng. EMG-Based Estimation of Limb Movement Using Deep Learning With Recurrent Convolutional Neural Networks. *Artificial Organs*, 42(5):E67–E77, 2018.

[Sup51] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[Sup52] Hao Zhu, Haotian Yang, Longwei Guo, Yidi Zhang, Yanru Wang, Mingkai Huang, Qiu Wu, Menghua and Shen, Ruigang Yang, and Xun Cao. FaceScape: 3D facial dataset and benchmark for single-view 3D face reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[Sup53] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[Sup54] Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European conference on computer vision*, pages 250–269. Springer, 2022.