SEREP: Semantic Facial Expression Representation for Robust In-the-Wild Capture and Retargeting

Arthur Josi [1,2*] Luiz Gustavo Hafemann[2*] Abdallah Dib[2] Emeline Got[2] Rafael M. O. Cruz[1]
Marc-André Carbonneau[2]

Ecole de Technologie Supérieure[1] Ubisoft LaForge[2]

1. Implementation details

For training the semantic expression model, we use the following importance weight in eq. 1 of the main manuscript. $\lambda_{rec} = 1.0$, $\lambda_{cycle} = 1.0$, $\lambda_{delta} = 0.01$, $\lambda_{edge} = 10000$ and $\lambda_{eyes} = 0.01$. Details for the model's architecture are provided Tab. 1.

For the expression capture model, the following importance weight are used in eq. 2 of the main manuscript: $\lambda_{code} = 10$, $\lambda_{lmks} = 1$ and $\lambda_{domain} = 0.005$. For the gradient reversal, we use a scale factor equal to 1. Details for the model's architecture are provided Tab. 2.

2. MultiREX additional details

We built the MultiREX benchmark from 8 identities from the MultiFace dataset [4]. We selected a single Range-of-Motion (ROM) sequence per identity, where the subjects perform a large variety of facial movements, including extreme expressions. For each ROM video, we consider 5 camera views, with the exception of identity '002914589' which only includes 4 (due to a camera failure). In total, we use 39 distinct videos. The benchmark comprises 10k ground truth meshes and 49k images. We note that while the original Multiface dataset contains 13 identities, we did not consider 5 subjects that either: (i) did not contain the range-of-motion sequence, (ii) had a camera failure for the frontal video or (iii) had videos that cropped a large portion of the subject's face.

Fig. 1 presents a frame for the different views used for evaluation, using the frame we manually selected for neutral representation of each individual. This is followed by the corresponding ground-truth mesh under the multiface topology and finally by the wrapped equivalent under the FLAME topology. FLAME neutrals are obtained using commercial software (Wrap 3D[1]), by first aligning each multiface mesh to the FLAME basehead with a rigid alignment with manually selected keypoints around eyes, nose, and mouth, then wrapping the mesh for topology conversion.

Figure 1. The 5 camera views used in MultiREX (left-to-right), followed by the corresponding ground-truth mesh under the Multiface and FLAME topology. We show all 8 subjects in the benchmark.

*Equal contribution

[1] https://faceform.com/

Function	Details
$\mathbf{E}_{\text{exp}}$	• **SpiralConv (x5)** `(3, 32)` → `(32, 32)` → `(32, 32)` → `(32, 64)` → `(64, 64)` • **Linear:** `Linear(3392, 64)`
$\mathbf{E}_{\text{id}}$	• **SpiralConv (x5)** `(3, 32)` → `(32, 32)` → `(32, 32)` → `(32, 64)` → `(64, 64)` • **Linear:** `Linear(3392, 64)`
$\mathbf{D}_{\text{mesh}}$	• **Linear:** `Linear(128, 3392)` • **SpiralConv (x5)** `(64, 64)` → `(64, 32)` → `(32, 32)` → `(32, 32)` → `(32, 3)`

Table 1. Model architecture for the semantic expression model.

Function	Details
$\mathbf{E}_{\text{img}}$	• **ConvNeXt-B()**
$\mathbf{H}_{\text{code}}$	• **ResBlock (x3):** `Linear(512, 512)` → `GELU` → `GroupNorm(32, 512)` • **Linear:** `Linear(512, 64)`
$\mathbf{H}_{\text{lmks}}$	• **Linear:** `Linear(512, 128)`
$\mathbf{C}_{\text{d}}$	• **GradientReversal()** • **Linear:** `Linear(512, 256)` → `GroupNorm(16, 256)` → `GELU` • **ResBlock (x2):** `Linear(256, 256)` → `GELU` → `GroupNorm(16, 256)` • **Linear:** `Linear(256, 1)`

Table 2. Model architecture for the expression capture model.

3. Ablations on the semantic expression model

We ablate the eye closure loss L_{eyes} and edge loss L_{edge} in Fig. 2. Without L_{eyes}, eye closure is incomplete during blinks. The edge loss follows existing practice [1] and improves animations' edge flow (removing jagged lines) to better support downstream animator needs. We will include these figures in the supplementary material.

The evaluation is inspired by the REALY benchmark [2]. Four masks are considered in the Multiface topology, as illustrated in Fig. 3. For a given mesh under evaluation, we first perform a rigid alignment of the evaluated regions from the ground truth to the generated mesh. For cheek and mouth, the rigid alignment is done with the combination of the mouth and cheek mask. After rigid alignment, we compute the mean vertex distance between the ground truth and the alignment mesh part.

4. More qualitative results

4.1. Comparison on MultiREX benchmark

In this section, we show more visual comparison against state-of-the-art methods on the MultiREX benchmark. Results are reported in Fig. 4, Fig. 5 and Fig. 6 for different subjects under different viewing angles. Our method shows robustness against side view changes and preserves better the subject's expression compared to other methods that generate less consistent expression over different views.

4.2. Comparison on in-the-wild images

Fig. 7, we show more qualitative results for in-the-wild retargeting to other subjects and considering different source

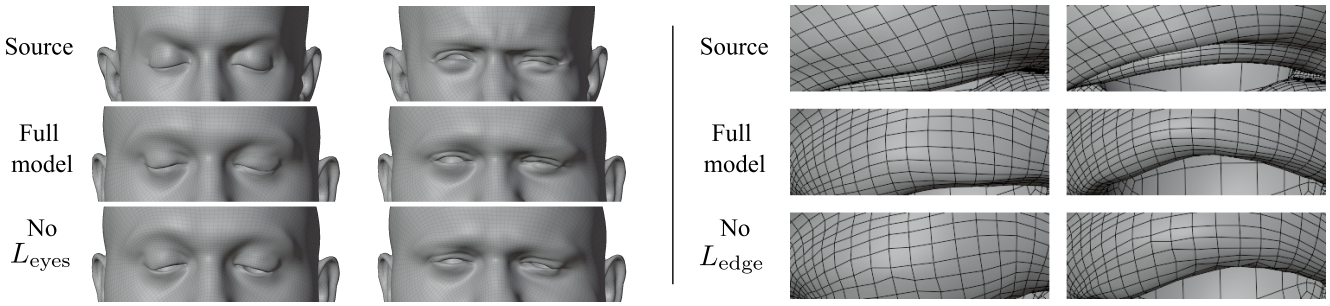

Figure 2. Ablation on L_{eyes} and L_{edge}.

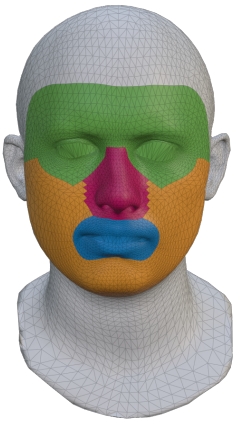

Figure 3. Visual representation of the forehead, nose, mouth, and cheek region masks used for our part-based evaluation. Masks do not overlap from one to another.

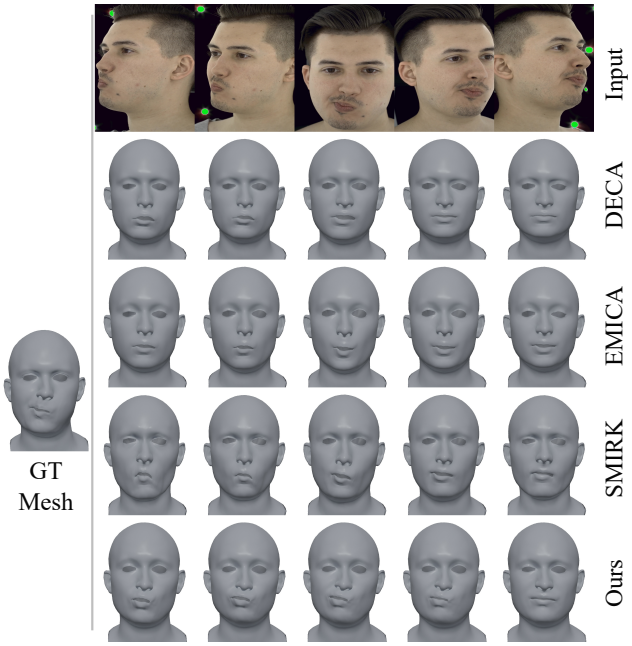

Figure 4. Additional captures on the proposed MultiREX benchmark

expressions. Our model is more faithful to the source expression in most scenarios while accounting for the face morphology and through its semantic expression model.

4.3. Challenging in-the-wild capture

We complete the evaluation of our expression capture model with challenging in-the-wild captures and some failure cases Fig. 8. Our model shows some robustness to lighting conditions, but can fail in cases of external occlusions, and motion blur. Motion blur is particularly problematic in video frames, resulting in jittery capture. The overall robustness of our method is in part dependent on the quality of the real landmarks used for training, obtained using an off-the-shelf detector.

5. Real-time processing

On a 2080Ti GPU (batch size of 1), the expression encoder takes 20.6 ms on average, and the mesh decoder takes 5.5 ms, for a total of 26.1 ms (38.3 FPS). This does not include bounding box detection.

6. Synthetic dataset

Fig. 9 shows samples from our synthetic dataset used to train our expression capture model. Rendering is done using Blender, with the Cycles renderer. We use randomly selected environment maps[2]. We use the same meshes and textures for teeth and eyes for all subjects. They are placed procedurally based on the vertex positions of the eyelids and jaw.

We emphasize that the generation of our synthetic dataset does not require 3D modeling for hair, facial accessories, or clothes, contrary to [3].

References

[1] Timo Bolkart, Tianye Li, and Michael J. Black. Instant multi-view head capture through learnable registration. In *Confer-*

[2]https://polyhaven.com

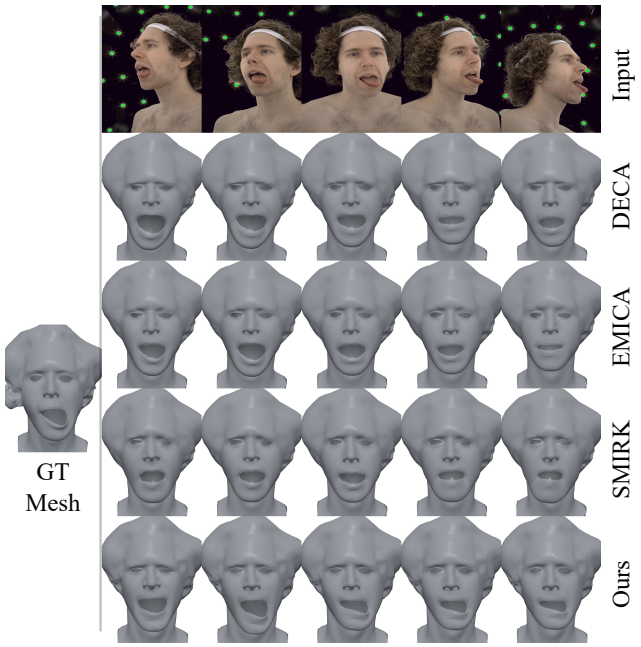

Figure 5. Additional captures on the proposed MultiREX benchmark

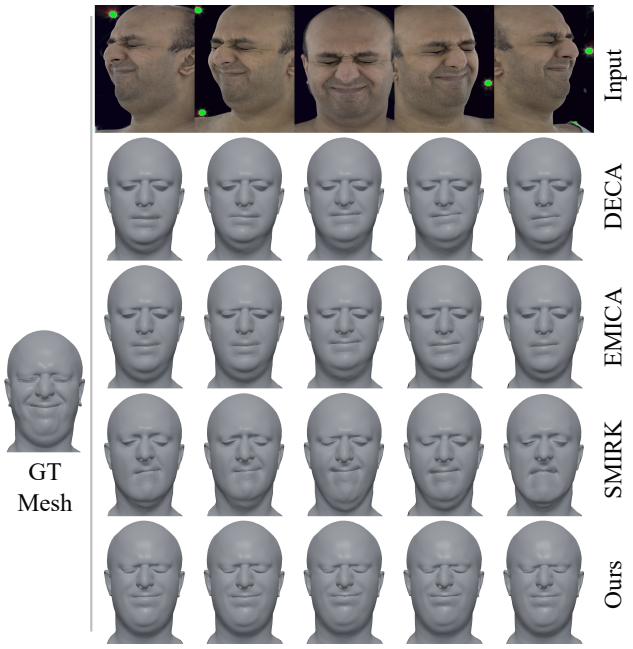

Figure 6. Additional captures on the proposed MultiREX benchmark

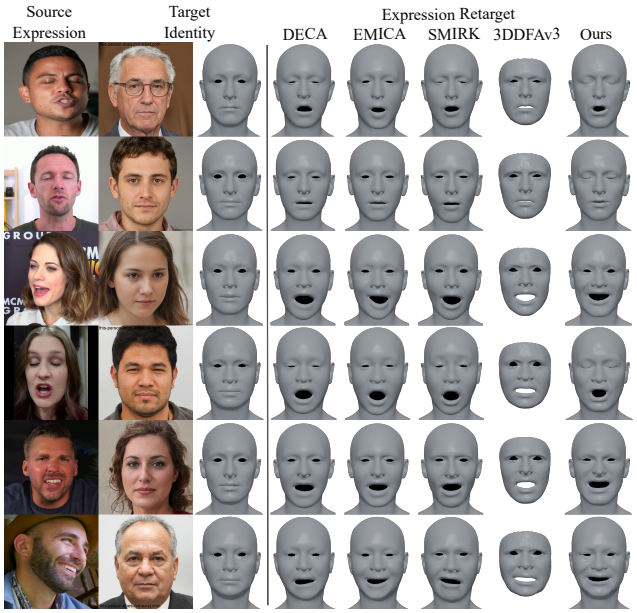

Figure 7. Additional retargeting on in-the-wild images

Figure 8. Results on challenging conditions

ence on Computer Vision and Pattern Recognition (CVPR), pages 768–779, 2023. 2

[2] Zenghao Chai, Haoxian Zhang, Jing Ren, Di Kang, Zhengzhuo Xu, Xuefei Zhe, Chun Yuan, and Linchao Bao. Realy: Rethinking the evaluation of 3d face reconstruction. In *European conference on computer vision*, pages 74–92. Springer, 2022. 2

[3] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3681–3691, 2021. 3

[4] Cheng-hsin Wuu, Ningyuan Zheng, Scott Ardisson, Rohan Bali, Danielle Belko, Eric Brockmeyer, Lucas Evans, Timothy Godisart, Hyowon Ha, Xuhua Huang, Alexander Hypes, Taylor Koska, Steven Krenn, Stephen Lombardi, Xiaomin

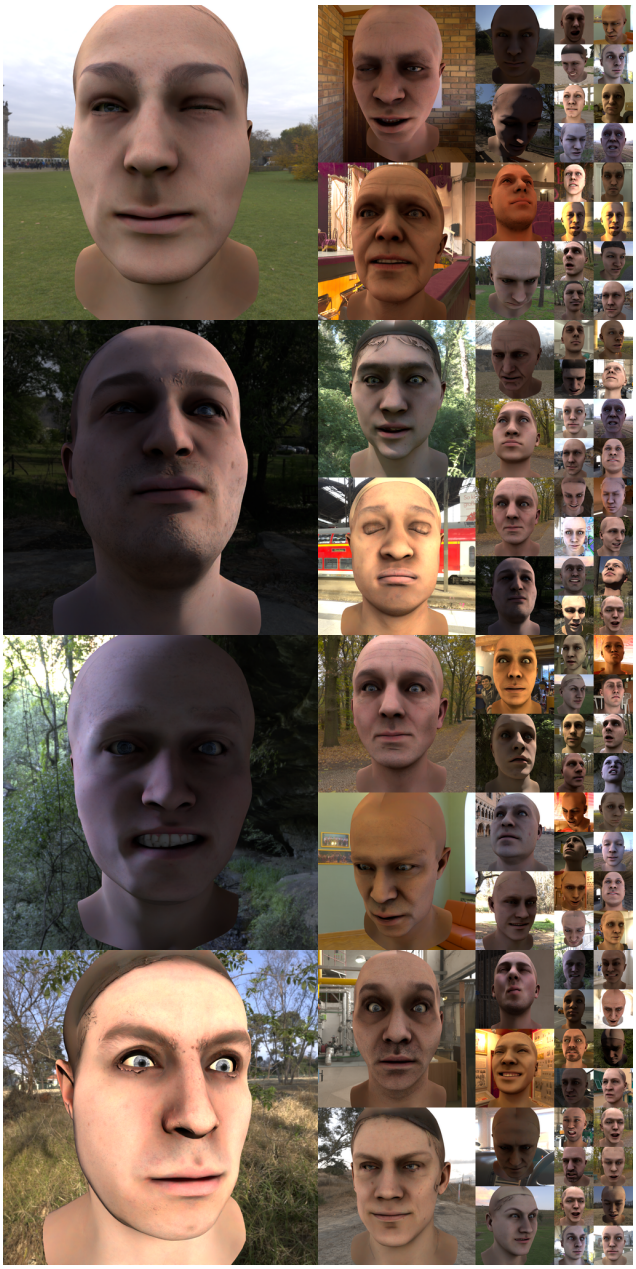

Figure 9. Random synthetic data samples.

Luo, Kevyn McPhail, Laura Millerschoen, Michal Perdoch, Mark Pitts, Alexander Richard, Jason Saragih, Junko Saragih, Takaaki Shiratori, Tomas Simon, Matt Stewart, Autumn Trimble, Xinshuo Weng, David Whitewolf, Chenglei Wu, Shoou-I Yu, and Yaser Sheikh. Multiface: A dataset for neural face rendering. In *arXiv*, 2022. 1