

SPARK: Self-supervised Personalized Real-time Monocular Face Capture

KELIAN BAERT, Technicolor Group, France and University Rennes, France
SHRISHA BHARADWAJ, Max Planck Institute for Intelligent Systems, Germany
FABIEN CASTAN, Technicolor Group, France
BENOIT MAUJEAN, Technicolor Group, France
MARC CHRISTIE, University Rennes, IRISA, CNRS, Inria, France
VICTORIA F. ABREVAYA, Max Planck Institute for Intelligent Systems, Germany
ADNANE BOUKHAYMA, Inria, University Rennes, IRISA, CNRS, France



Fig. 1. After a personalization step, SPARK can reconstruct accurate faces with expressive details on unseen images **in real-time**.

Feedforward monocular face capture methods seek to reconstruct posed faces from a single image of a person. Current state of the art approaches have the ability to regress parametric 3D face models in real-time across a wide range of identities, lighting conditions and poses by leveraging large image datasets of human faces. These methods however suffer from clear limitations in that the underlying parametric face model only provides a coarse estimation of the face shape, thereby limiting their practical applicability in tasks that require precise 3D reconstruction (aging, face swapping, digital make-up,...). In this paper, we propose a method for high-precision 3D face capture taking advantage of a collection of unconstrained videos of a subject as prior information. Our proposal builds on a two stage approach. We start with the reconstruction of a detailed 3D face avatar of the person, capturing both precise geometry and appearance from a collection of videos. We then use the encoder from a pre-trained monocular face reconstruction method, substituting its decoder with our personalized model, and proceed with transfer learning on the video collection. Using our pre-estimated image formation model, we obtain a more precise self-supervision objective, enabling improved expression and pose alignment. This results in a trained encoder capable of efficiently regressing pose and expression parameters in real-time from previously unseen images, which combined with our personalized geometry model yields more accurate and high fidelity mesh inference.

Through extensive qualitative and quantitative evaluation, we showcase the superiority of our final model as compared to state-of-the-art baselines, and demonstrate its generalization ability to unseen pose, expression and lighting.

CCS Concepts: • Computing methodologies → Machine learning.

Additional Key Words and Phrases: Face Capture, Face Reconstruction, Personalized Avatars

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24)*, December 3–6, 2024, Tokyo, Japan, <https://doi.org/10.1145/3680528.3687704>.

ACM Reference Format:

Kelian Baert, Shrisha Bharadwaj, Fabien Castan, Benoit Maujean, Marc Christie, Victoria F. Abrevaya, and Adnane Boukhayma. 2024. SPARK: Self-supervised Personalized Real-time Monocular Face Capture. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24)*, December 3–6, 2024, Tokyo, Japan. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3680528.3687704>

1 INTRODUCTION

3D facial performance capture is a key component in several applications, including immersive telepresence in AR or VR and visual effects for the entertainment industry. Producing high-quality results, however, often requires large financial, time and resource investments. This involves not only expensive 3D capture equipment [Beeler et al. 2010; Debevec et al. 2000], precise marker-based tracking systems [Bennett and Carter 2014] or head-mounted displays [Brito and Mitchell 2019], but also extensive hours of capture sessions from the actor. Marker-less capture setups are promising solutions for simplifying this pipeline, however high-quality results still rely on complex rigs [Helman et al. 2020] or large, personalized training datasets [Laine et al. 2017; Wu et al. 2018].

At the other end of the spectrum are 3D reconstruction methods that can operate on images or videos from affordable consumer-grade hardware. The main idea is to use statistical models of the 3D face – so-called 3D Morphable Models (3DMMs) – which are fitted to RGB images or 2D landmarks using optimization-based [Andrus et al. 2020; Zielonka et al. 2022] or learning-based methods [Danecek et al. 2022; Feng et al. 2021; Retsinas et al. 2024]. The prior knowledge from the statistical model helps to overcome the ill-posed nature of the problem, and the development of learning-based techniques has enabled unprecedented robustness to pose, illumination and occlusions. However, this comes at the cost of lower geometric quality, providing only a coarse approximation of the shape and expression which falls well short of high-end systems.

Neural head avatars were recently proposed to address some of the limitations of 3DMM-based methods. The goal here is to train a model from a single monocular video, that can afterwards be animated with novel expressions [Grassal et al. 2022; Zheng et al. 2022, 2023] and optionally environments [Bharadwaj et al. 2023]. The key idea is to harness advances in neural rendering [Hasselgren et al. 2022; Munkberg et al. 2022] which allow to better model geometric and appearance details. While this results in images of higher quality, the estimated 3D geometry still remains behind production systems. Furthermore, the computational costs of such approaches prevent them from running in real-time or even close to real-time contexts.

In this paper we aim to bridge the gap between these two worlds by answering the following: **given a collection of unconstrained videos of a person as prior information, can we build a personalized real-time tracker that regresses higher quality geometry?** Video collections offer a valuable middle ground between high-end capture systems and consumer-grade solutions. Without requiring complex, non-portable hardware, we can leverage multiple videos with a variety of illuminations and potentially more diverse poses to better constrain the problem of 3D face reconstruction.

More specifically, we propose SPARK, a novel method to build *personalized*, high-quality 3D facial models that can be tracked in real-time. Our approach consists of two key steps: (1) constructing a personalized 3D mesh and deformation model from a collection of *in-the-wild* videos, and (2) training a fast, regression-based tracker that can provide accurate reconstructions given a novel image. The challenge here is two-fold. First, we need to design a system with the capacity to distill high-fidelity geometry from contents with very diverse lighting conditions, color tones and appearances. To this end, we leverage recent advances in unsupervised efficient inverse neural rendering (*e.g.* [Bharadwaj et al. 2023; Zheng et al. 2022; Zielonka et al. 2023]) and adapt them to the multi-video case. Second, we need to recover a high fidelity topologically consistent 3D facial geometry in real time. Rooted in transfer learning, our idea is to consolidate the benefits of state-of-the-art generalizable 3D face capture methods (*e.g.* [Danecek et al. 2022; Feng et al. 2021]) that were trained on a large *in-the-wild* image corpora, and propose an efficient method for adapting to the specific subject, leveraging the results from the previous step.

Through quantitative and qualitative evaluations, our experimental results show that we can generalize successfully to unseen lighting, expressions and poses, and are capable of producing accurate geometries in real time. Our method outperforms competing techniques with similar inference times offering, more accurate poses and expressions, together with more expressive meshes.

In summary, our contributions are:

- A new method using multiple monocular videos of a person to estimate the 3D shape and range of deformations of a face, with fine geometric details essential for facial expressiveness.
- A transfer learning approach that enables fine-tuning a face tracking method for a specific subject using a pre-estimated personalized geometry and appearance model, while still being able to generalize to different lightings, poses and expressions.

- Thorough evaluations of our proposed approach, including two new metrics for evaluating the posed geometry without 3D ground truth.

2 RELATED WORK

Head avatars from video. There have been many efforts towards the democratization of animatable 3D head avatar creation via commodity sensors, as opposed to costly traditional capture systems [Beeler et al. 2011; Debevec et al. 2000; Ghosh et al. 2011; Riviere et al. 2020]. Traditional methods [Garrido et al. 2016; Thies et al. 2016a] utilize statistical models [Blanz and Vetter 1999] to reconstruct 3D shape and appearance, but they result in relatively coarse reconstructions. NerFACE [Gafni et al. 2021] pioneered in using dynamic neural radiance fields (NeRF) [Mildenhall et al. 2020] to create head avatars. IMavatar [Zheng et al. 2022] achieves precise geometry recovery by using implicit surfaces, simultaneously learning canonical head geometry and expression deformations. However, methods relying on implicit representations often suffer from inefficiencies in training and rendering. PointAvatar [Zheng et al. 2023] adopts a similar deformation model but uses an explicit point cloud representation, facilitating faster rasterization and improved image quality. Recently, several approaches [Gao et al. 2022; Xu et al. 2023; Zielonka et al. 2023] have leveraged InstantNGP [Müller et al. 2022] to accelerate radiance field queries, enabling avatar reconstruction within 5 to 20 minutes. Contemporary work proposes to build avatars through Gaussian Splatting (3DGS) [Kerbl et al. 2023] by anchoring gaussians to a coarse underlying control geometry [Ma et al. 2024; Qian et al. 2023; Xu et al. 2024]. However, there is no clearly established mechanism to robustly extract topologically consistent detailed meshes from these models. Neural Head Avatar [Grassal et al. 2022] creates mesh-based avatars containing full head and hair geometry. Despite this, the resulting geometry is relatively coarse, with many details rendered in the texture space. The recent FLARE [Bharadwaj et al. 2023] constructs high-quality mesh-based avatars within 15 minutes, while decomposing appearance into albedo, roughness, and extrinsic illumination. We build on this method in our avatar reconstruction *in-the-wild* stage, by extending it to a novel multi monocular video setting, in addition to introducing several improvements to enhance efficiency and expressiveness.

3D monocular face capture. Model-free methods can learn to directly infer meshes [Abrevaya et al. 2019; Alp Guler et al. 2017; Deng et al. 2020; Dou et al. 2017; Feng et al. 2018; Jung et al. 2021; Ruan et al. 2021; Sela et al. 2017; Szabó et al. 2019; Wei et al. 2019; Wu et al. 2020; Zeng et al. 2019], voxels [Jackson et al. 2017], or Signed Distance Functions [Yenamandra et al. 2021]. However, they require extensive 3D training data, and they can also suffer from limited expressiveness and generalization due to the synthetic/real domain gap [Dou et al. 2017; Sela et al. 2017; Zeng et al. 2019], with many of them relying on synthetic training data. 3DMMs (*e.g.* BFM [Paysan et al. 2009], FaceWarehouse [Cao et al. 2013], FLAME [Li et al. 2017]) can be fitted to images through test-time analysis-by-synthesis optimization [Aldrian and Smith 2012; Bas et al. 2017; Blanz et al. 2002; Blanz and Vetter 1999; Cao et al. 2014; Garrido et al. 2016; Gerig et al. 2018; Koizumi and Smith 2020; Li et al. 2013;

Ploumpis et al. 2021; Romdhani and Vetter 2005; Thies et al. 2015, 2016a,b], but these are not useful for time-sensitive applications. Recent deep learning based models offer robust and fast inference, either in their supervised [Cao et al. 2015; Chang et al. 2018; Guo et al. 2020; Kim et al. 2018; Richardson et al. 2016; Tran et al. 2017, 2018; Zhang et al. 2023; Zhu et al. 2016; Zelionka et al. 2022] or weakly/self-supervised [Abrevaya et al. 2020; Alp Guler et al. 2017; Danecek et al. 2022; Deng et al. 2019; Feng et al. 2021; Genova et al. 2018; Liu et al. 2017; Sanyal et al. 2019; Shang et al. 2020; Tewari et al. 2019, 2018, 2017; Wood et al. 2022; Yang et al. 2020] forms. Most notably, recent self-supervised models DECA [Feng et al. 2021] and EMOCA [Danecek et al. 2022] offer arguably state-of-the-art animatable geometry inference performance. Their encoders predict 3DMM, camera and spherical harmonics coefficients used to render and compare a morphable geometry to images in the wild for large scale training. Despite their additional mesh detail regression, they still showcase expression generalization and alignment issues, and their generic FLAME [Li et al. 2017] based geometry can lack expressiveness and fidelity. To remedy these issues, and in contrast with existing literature, we propose to personalize such models using solely a few monocular videos.

3 METHOD

At inference, our system takes as input an image of a subject and generates a 3D face reconstruction including pose and expression in real time. At training time, our work follows a two-stage approach as illustrated in Fig. 2. First we leverage an unconstrained collection of video sequences of the same subject to build a personalized decoder using inverse rendering in a stage called MultiFLARE, in which we disentangle the color into illumination and intrinsic material (see top row in Fig. 2). Second, we adapt a generalizable feedforward 3D face capture model by replacing its decoder with the custom MultiFLARE geometry model, and tuning its encoder on the video collection using the pre-computed reflectance functions. In the following, we detail these two stages.

3.1 MultiFLARE: Inverse rendering from a video collection

Feedforward face reconstruction methods leverage a generic statistical model for geometry and appearance, serving as a differentiable image formation layer for image based self-supervised learning. We replace this with a new differentiable decoder built efficiently from our adaptation data through inverse rendering. Conjointly, the inverse rendering allows us to estimate a reflectance function per adaptation sequence. We detail in the following the optimization leading to these personalized geometry and reflectance functions.

Geometry and Deformation. We use a triangular mesh as our geometry representation and build on the FLAME 3DMM [Li et al. 2017]. The personalized shape is obtained in canonical space by optimizing the template vertices, similarly to PointAvatar [Zheng et al. 2023] and FLARE [Bharadwaj et al. 2023]. A given point x_c in canonical space is then deformed for pose and expression using the following equation:

$$\begin{aligned} \text{FLAME}(x_c, \mathcal{P}, \mathcal{E}, \mathcal{W}, \theta, \psi) = \\ \text{LBS}(x_c + B_P(\theta; \mathcal{P}) + B_E(\psi; \mathcal{E}), J(\beta), \theta, \mathcal{W}) \end{aligned} \quad (1)$$

where LBS is the standard linear blend-skinnning function, J is the joint regressor used to compute joint locations from mesh vertices, $\mathcal{W} \in \mathbb{R}^{nv \times n_j}$ are per-vertex blend-skinnning weights, $B_P(\cdot)$ and $B_E(\cdot)$ compute pose and expression offsets given pose correctives \mathcal{P} and expression blendshapes \mathcal{E} . β , θ and ψ are respectively pre-computed FLAME parameters for shape, pose and expression. We follow IMAvatar [Zheng et al. 2022] and FLARE in personalizing the expression blendshapes by training a deformation network \mathcal{D} that predicts the expression basis $\mathcal{E} \in \mathbb{R}^{ne}$ given canonical vertex location x_c :

$$\mathcal{D}(\gamma(x_c)) : \mathbb{R}^3 \rightarrow \mathcal{E} \quad (2)$$

We apply sinusoidal positional encoding $\gamma(\cdot)$ to map canonical positions into a higher dimensional space, enabling the network to model higher frequency variations. \mathcal{D} is initialized in a separate supervised pre-training stage that minimizes $\|\mathcal{D}(\gamma(x)) - \mathcal{E}\|_2$ at the positions of the canonical FLAME vertices, such that the network initially mimics the FLAME expression basis \mathcal{E} . Note that this stage is completed in under a minute, as we are only performing forward and backward passes through \mathcal{D} and not rendering.

We perform remeshing [Botsch and Kobbelt 2004] of the canonical geometry after a fixed number of training iterations to increase its resolution.

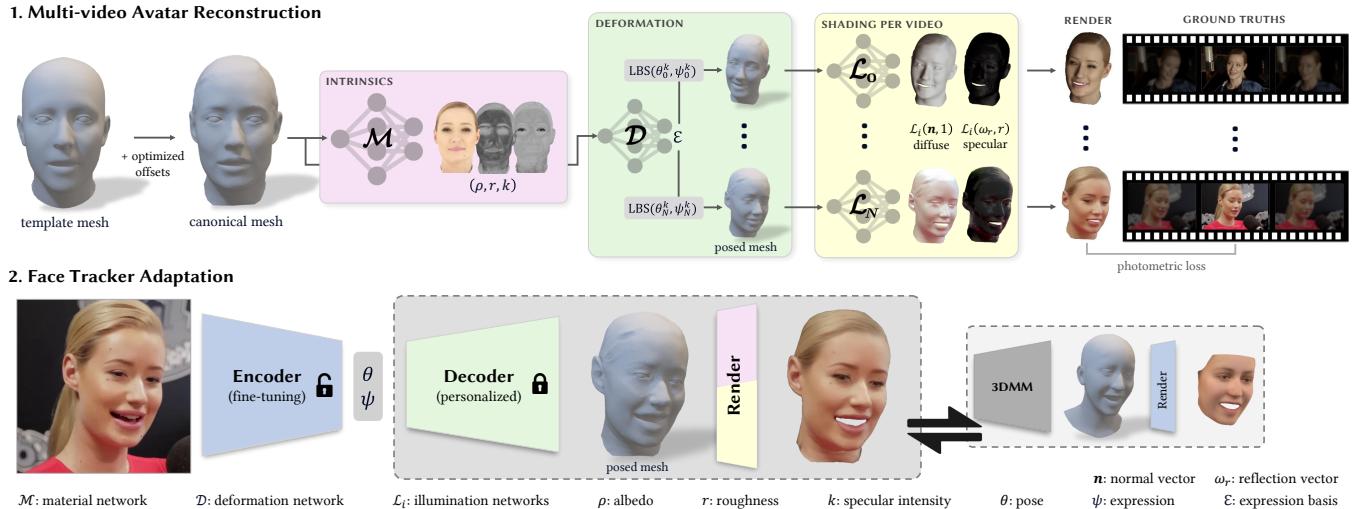
Illumination. Modelling the appearance is challenging in our setting as we have N videos of each subject with varying illumination conditions. Moreover, the personalized canonical geometry is dependent on the learning of the appearance, as we optimize the geometry through deferred shading. Hence, to handle multiple videos, we disentangle the color into illumination and intrinsic materials such as albedo, roughness and specular intensity. However, our dataset by design has in-the-wild videos that are low-dynamic range and have uncontrolled lighting conditions such as over/under-exposed frames. Thus, it is non-trivial to model the illumination and disadvantageous to represent it using HDRI maps. To tackle this, we follow FLARE and split the rendering equation into a diffuse term using a Lambertian model and a specular term through the Cook-Torrance microfacet model [Cook and Torrance 1982], which is evaluated using the neural split-sum approximation. More specifically, the illumination is represented by a single MLP controlled using Integrated Directional Encoding [Verbin et al. 2022] (IDE) to mimic different mipmap levels regulated by the surface roughness. We kindly refer the readers to FLARE [Bharadwaj et al. 2023] for more details and for the sake of simplicity, we denote the neural split-sum approximation (including IDE and the precomputed look-up table *FG-LUT*) as \mathcal{L} , such that:

$$\mathcal{L}(n, 1) = l_d \quad \mathcal{L}(\omega, r) = l_s \quad (3)$$

where n is the surface normal, ω is the reflection vector, r is the roughness, l_d represents the diffuse shading and l_s represents the specular shading. To adapt to our setting, we extend this to N approximations given by $\{\mathcal{L}^1, \dots, \mathcal{L}^N\}$ and optimize the parameters for each illumination separately: $\{l_d^1, \dots, l_d^N\}$ and $\{l_s^1, \dots, l_s^N\}$.

Intrinsic material properties. We learn the albedo, roughness and specular intensity through a MLP \mathcal{M} in canonical space as introduced by FLARE. Contrary to the illumination, which is learned

1. Multi-video Avatar Reconstruction



2. Face Tracker Adaptation

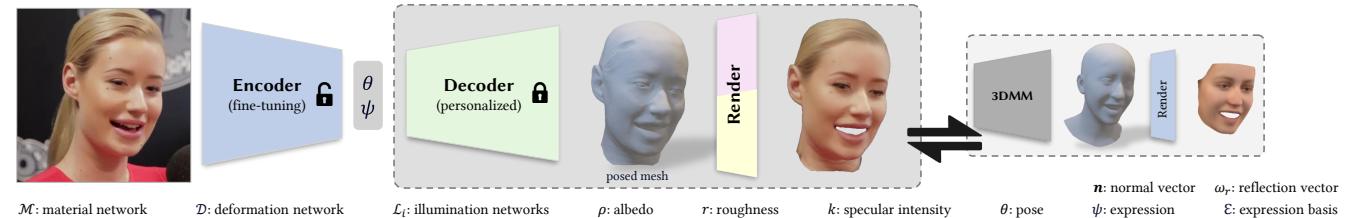


Fig. 2. Illustration of our two-stage adaptation process. In stage 1, we rely on a collection of different video sources of the same person to build a personalized geometry decoder through inverse rendering. In stage 2, the 3DMM of a generalizable feedforward face capture network is swapped with the new decoder, and the encoder is tuned by reconstructing the same adaptation video frames leveraging the pre-estimated reflectance function for each video.

separately for each video, we make the assumption that the person's intrinsic appearance varies little between videos and learn a single set of material properties. \mathcal{M} takes canonical points (x_c) as inputs and predicts the albedo ρ , roughness r and specular intensity k :

$$\mathcal{M}(x_c) : \mathbb{R}^3 \rightarrow \rho, r, k \quad (4)$$

The final color for the i^{th} video is calculated as follows:

$$C = \rho \cdot l_d^i + k \cdot l_s^i \quad (5)$$

Thus, we enforce consistency in the intrinsic materials by design by using the same network \mathcal{M} to calculate the final color for N illuminations. The learnable illumination, along with shared subject-specific geometry and material properties are optimized through inverse rendering using differentiable rasterization.

3.2 Face Tracker Adaptation

Standard 3D face reconstruction networks (e.g. [Danecek et al. 2022; Feng et al. 2021; Retsinas et al. 2024]) typically consist of an encoder, that we denote E_ϕ , with parameters ϕ . They predict shape, pose and expression parameters, among others. Combined with a 3D morphable model and a relatively simple appearance representation, such methods learn self-supervisedly from images through inverse rendering. We wish to benefit from the feedforward prediction capabilities of such encoders, whose large-scale training enables generalization to a large variety of illuminations, poses and expressions, and associate this ability with our superior expressive personalized geometry decoder, as opposed to a vanilla 3DMM. Without loss of generality, we build on EMOCA [Danecek et al. 2022], which uses the FLAME 3DMM.

From the previous stage, we recover a personalized geometric model, through the identity-specific canonical positions x_c and expression deformation basis \mathcal{E} (Eq. 1). We also obtain a personalized

reflectance function (Eq. 5) allowing us to render this custom geometry, with intrinsic materials of the subject's face (ρ, r, k), along with lighting for the training sequences (l_d^i, l_s^i). We now freeze our canonical geometry, appearance and deformation models. We compute the expression basis \mathcal{E} and intrinsic materials (ρ, r, k) according to Eq. 2 and Eq. 4 respectively. Whilst our personalized MultiFLARE geometry model is built starting from FLAME, its final canonical geometry and deformation basis deviate from their original definition due to their personalization (Section 3.1). In fact, simply swapping FLAME for our personalized estimations in EMOCA leads to very suboptimal results, as can be seen in Table 2.

Hence, inspired by few-shot adaptation literature (e.g. [Gao et al. 2024; Zhou et al. 2022]), we propose to tune the encoder in order to adapt its output latent space (θ and ψ) to our new personalized geometry representation. Specifically, we update the weights of the last ResNet block [He et al. 2016] of the backbone and the entire MLP head, for both the coarse shape encoder and the expression encoder of EMOCA. This allows us to adapt to the modified regression task while retaining the general features learned in earlier layers [Lee et al. 2022], as our objective can be akin to alleviating output level shift in the regression task.

3.3 Training details

In this section, we provide information regarding the training of both stages of our method. More details are available in our supplemental material.

3.3.1 MultiFLARE

FLAME regularization. Previous methods, namely the aforementioned IMavatar, PointAvatar and FLARE, regularize the deformations using the FLAME values at the nearest vertex, which can be computed for arbitrary mesh topologies and various geometry

representations. Our original canonical geometry uses the FLAME topology and only changes at the remeshing step. Thus, we update the FLAME expression basis by projecting the remeshed vertex positions on the original mesh and using barycentric interpolation. We then use those updated per-vertex values directly in place of the nearest neighbors, keeping the same loss formulation. We find this improves stability for learning the deformation network, especially for identities whose canonical geometry deviates heavily from the FLAME template.

Objective function. Our full objective function is as follows:

$$\begin{aligned} L = & \lambda_{\text{RGB}} L_{\text{RGB}} + \lambda_{\text{vgg}} L_{\text{vgg}} + \lambda_{\text{mask}} L_{\text{mask}} + \lambda_{\text{FLAME}} L_{\text{FLAME}} + \\ & \lambda_{\text{laplacian}} L_{\text{laplacian}} + \lambda_{\text{normal}} L_{\text{normal}} + \lambda_{\text{smooth}} L_{\text{smooth}} + \\ & \lambda_r L_r + \lambda_{\text{spec}} L_{\text{spec}} + \lambda_{\text{light}} L_{\text{light}} \end{aligned} \quad (6)$$

where L_{RGB} and L_{vgg} are respectively a photometric L2 loss in log space and a perceptual loss [Simonyan and Zisserman 2015] between masked ground-truth and rendered image, L_{mask} is an L2 loss between ground truth and rasterized binary mask, $L_{\text{laplacian}}$ is a laplacian smoothness regularizer for the canonical vertices, L_{normal} is a cosine similarity loss between neighboring face normals of the canonical geometry. L_{FLAME} regularizes the expression blendshapes \mathcal{E} from the deformation network \mathcal{D} using the FLAME basis, L_r and L_{spec} enforce a predefined distribution for specular intensity k and surface roughness r , L_{light} penalizes strong deviations from white light in the diffuse shading and L_{smooth} is a smoothness regularization for albedo and roughness. We refer the reader to FLARE for more details on the individual loss functions. We set the relative weights λ_i of those terms as follows: $\lambda_{\text{RGB}} = 1.0$, $\lambda_{\text{vgg}} = 0.1$, $\lambda_{\text{mask}} = 2.0$, $\lambda_{\text{FLAME}} = 20.0$, $\lambda_{\text{laplacian}} = 100.0$, $\lambda_{\text{normal}} = 0.1$, $\lambda_{\text{smooth}} = \lambda_r = \lambda_{\text{spec}} = \lambda_{\text{light}} = 0.01$.

Single-stage training. We jointly optimize the canonical positions x_c , deformation \mathcal{D} , material \mathcal{M} and lighting networks (l_d^i, l_s^i) in a single stage, with the material network equipped from the start with hash-grid encoding [Müller et al. 2022]. To address the findings of FLARE, wherein the model would overfit to color too quickly resulting in over-smooth shapes, we use *progressive* multi-resolution hash encoding [Li et al. 2023] and enable the increasingly high-resolution hash levels successively. Through this change, we achieve high quality texture and geometry without having to learn \mathcal{M} and (l_d^i, l_s^i) from scratch in a second stage. We exponentially decay the learning rate of the canonical vertices at a rate of $\alpha = 0.998$ per iteration. Additionally, we conjointly fine-tune the pre-estimated pose and expression parameters to allow for better alignment for frames where the pre-estimated values are imprecise.

3.3.2 Face Tracker Adaptation.

Objective function. The transfer learning stage described in Section 3.2 is performed through gradient descent using the following combined loss, borrowed from the coarse stage training of EMOCA v2 [Danecek et al. 2022; Filntisis et al. 2023], and applied to the

training video frames of the identity of interest:

$$\phi^* = \operatorname{argmin}_\phi L(\phi) \quad (7)$$

$$\begin{aligned} L(\phi) = & \lambda_{\text{emo}} L_{\text{emo}} + \lambda_{\text{pho}} L_{\text{pho}} + \lambda_{\text{lmk}} L_{\text{lmk}} + \lambda_{\text{eye}} L_{\text{eye}} + \\ & \lambda_{\text{mc}} L_{\text{mc}} + \lambda_\psi L_\psi + \lambda_{\text{lipr}} L_{\text{lipr}} \end{aligned} \quad (8)$$

where L_{emo} , L_{pho} , L_{lmk} , L_{eye} , L_{mc} , L_ψ are respectively the emotion consistency loss, photometric loss, landmarks reprojection loss, relative landmarks-based eye closure loss and mouth closure loss, and expression regularization loss as introduced in EMOCA. L_{lipr} is the lip reading loss based on [Filntisis et al. 2023] introduced in EMOCA v2. λ_i designates the respective weight of each objective. We remind that the self-supervised rendering loss L_{pho} uses our personalized geometry model (Eq. 1) in combination with our reflectance (Eq. 5), computed using the corresponding lighting l_d^i, l_s^i for each adaptation video.

3.4 Test-time inference

After the encoder adaptation, geometry prediction for any unseen image I of the target identity is obtained by combining the inferred parameters $(\theta^*, \psi^*) = E_{\phi^*}(I)$ with our personalized MultiFLARE geometry model (Eq. 1). As our neural geometry model is pre-computed after the personalization stage, we can perform inference in real time on standard GPUs (forward pass through E_{ϕ^*} , a Resnet-50 [He et al. 2016] network).

4 EXPERIMENTS

In this section we demonstrate our method on 6 datasets of different subjects, chosen for diversity and ease of access to in-the-wild videos. Each dataset is made up of 6 to 12 in-the-wild videos with durations ranging between 10 seconds and 1 minute. These videos are cut from interviews, talks and dialogue scenes. We provide sources for the videos in our supplemental material. For practicality purposes, we subsample all sequences at a 1:4 ratio for training. We train MultiFLARE at a resolution of 512×512 . The images are further cropped and resized to 224×224 for the transfer learning stage, a standard resolution used by many recent feedforward face reconstruction methods.

We compare our method with state-of-the-art feedforward face capture methods: DECA [Feng et al. 2021], EMOCA [Danecek et al. 2022] and SMIRK [Retsinas et al. 2024]. Additionally, for fair comparison, we evaluate against EMOCA fine-tuned, per identity, on the data used for building our adapted feedforward model.

4.1 Qualitative Evaluation

Multi-video avatar reconstruction. Given N videos, SPARK first reconstructs a head avatar using MultiFLARE. Fig. 9 shows examples of the reconstructed image, the albedo, shading and geometry for a few subjects. By leveraging multiple monocular sequences of a person, we are able to disentangle the intrinsic face albedo from the shading and recover fine geometric details. Additionally, our improved *deformation* network (Eq. 2) is able to model precise subtle expressive details.

Personalized real-time face capture. At test time, SPARK is able to reconstruct 3D geometry in real-time given unseen images of

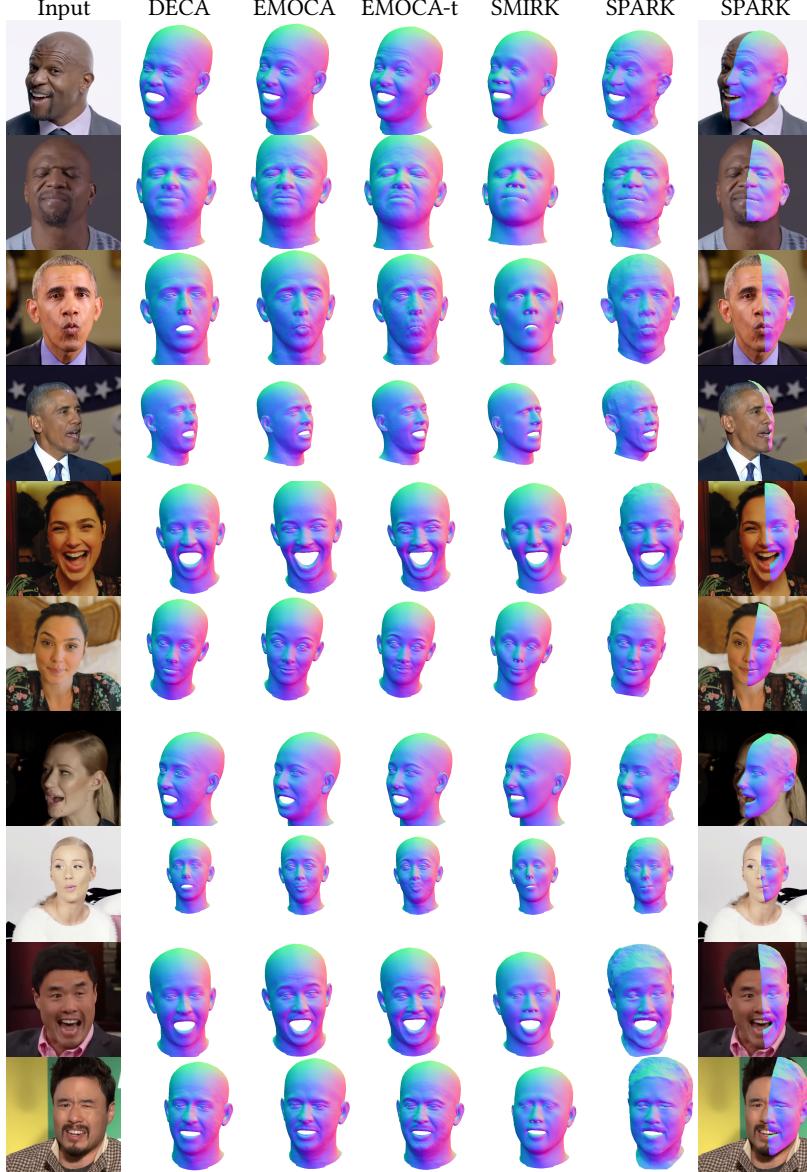


Fig. 3. Comparison to multiple state of the art face reconstruction methods. From left to right: input image, DECA [Feng et al. 2021], EMOCA [Danecek et al. 2022], EMOCA fine-tuned, SMIRK [Retsinas et al. 2024] and SPARK. See Fig. 8 for more examples. More results are also provided in our supplemental videos.

the person. Figure 3 shows qualitative results of SPARK compared against existing methods. Our approach is able to reconstruct faces that match the input image better than the previous state-of-the-art. In addition to more precise overall alignment of facial features, we find that our improved *deformation* model can represent fine dynamic details of the face such as wrinkles and cheek movement under large jaw poses. This demonstrates our method’s ability to reconstruct details beyond what 3DMMs such as FLAME can represent, even on unseen images of the person, under illuminations and poses that differ from the training set of videos. More results are available in the supplemental videos, where we also demonstrate

how our accurate tracking can be used for face editing tasks such as adding facial hair through texture mapping.

4.2 Quantitative Evaluation

Existing face capture methods generally do not evaluate the accuracy of the posed geometry. One common scheme is to evaluate the estimated neutral face using 3D ground truth data, with benchmarks such as NoW [Sanyal et al. 2019]. Differently, we seek to evaluate the quality of the estimated poses and expressions. However, to our knowledge there is no dataset that comprises multiple monocular in-the-wild videos of the same subject in different contexts with

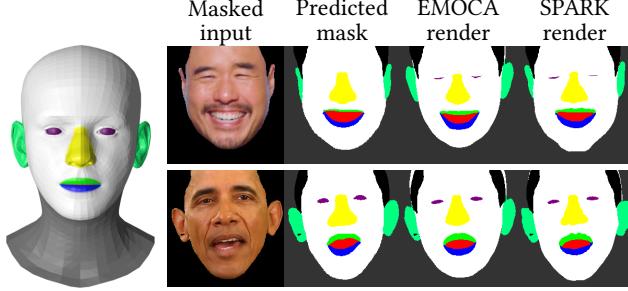


Fig. 4. Illustration of our semantic Intersection-over-Union metric. Left: manually annotated semantic masks for FLAME. Right: two examples with ground-truth segmentation, a render of EMOCA’s tracked mesh and our tracked mesh.

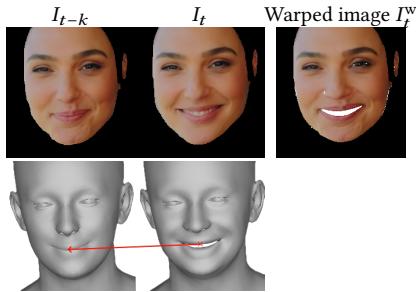


Fig. 5. Illustration of our image warping metric. The background and hair are masked out in both frames. We use the tracked geometry to backtrack each rasterized pixel of image I_t to a pixel in image I_{t-k} . The white area is the occlusion mask, where we do not compare pixels.

per-frame 3D ground truths for evaluation. We also argue that even recent dense landmark prediction models remain too sparse to accurately evaluate state of the art dense geometric face capture methods. Additionally, since landmarks are often used to train such methods, the specific landmarks layout chosen for evaluation could yield an unfair advantage to some methods over others (overfitting). Thus, we introduce two new metrics for evaluating the posed geometry independently of albedo and shading.

Semantic IoU metric. We use state-of-the-art semantic segmentation method BiSeNet [Yu et al. 2018] to predict semantic masks for the following head regions: skin, hair, nose, ears, eyes, upper lip, lower lip and mouth interior. We mask out the hair region, as hair is not modeled by current feedforward face capture methods. We manually annotate the FLAME template mesh with corresponding semantic masks, as shown in Fig. 4, allowing us to render a semantic segmentation from the posed geometry. We compute the mean intersection over union (IoU) over all the semantic classes of a given frame:

$$\text{IoU}_S = \frac{1}{C} \sum_{i=1}^C \frac{|M_i \cap \hat{M}_i|}{|M_i \cup \hat{M}_i|}$$

where C is the number of classes considered, \hat{M}_i are the respective pseudo ground-truth semantic masks and M_i are the rasterized

Method	Warp PSNR \uparrow	Semantic IoU \uparrow	Landmarks L1 \downarrow
DECA	29.501	0.616	0.047
SMIRK	30.062	0.649	0.035
EMOCA	29.979	0.644	0.036
EMOCA (fine-tuned)	30.275	0.673	0.027
Ours	30.647	0.702	0.027

Table 1. Quantitative results assessing the accuracy of the captured geometry on unseen data.

masks. In Fig. 4 we show examples of the predicted and rasterized masks. The semantic IoU error measures the alignment of various face regions independently of external factors such as scene illumination, motion blur and high frequency variation on the skin. However, it is sensitive to the accuracy of the semantic segmentation.

Geometry-based image warping metric. We introduce a second metric directly based on the input images to measure the accuracy of tracked geometry. The video sequence is sampled at fixed intervals k . For each pair of consecutively sampled frames (I_{t-k}, I_t), we use the tracked geometry to reconstruct image I_t from pixels of I_{t-k} . More precisely, we use rasterized barycentric coordinates to map each pixel in I_t to a surface point on the tracked geometry. We then project the corresponding point on the geometry of I_{t-k} to screen space and retrieve the pixel value, forming a new warped image I_t^w . The image warping error is defined as follows:

$$W_{\text{PSNR}}(t) = \text{PSNR}(I_t \cdot o_{t-k,t}, I_t^w \cdot o_{t-k,t})$$

where $o_{t-k,t}$ is a binary occlusion mask for surface points that were not visible in I_{t-k} . Note that we use semantic segmentation masks to remove the background and hair from both images. In Fig. 5, we illustrate the warping on a pair of images. While this metric can be computed for any arbitrary pair of images, we find that it is more reliable with a short frame interval, reducing potential issues such as aliasing, incorrect shading and large occlusion masks. We empirically choose an interval of 170 ms (5 frames at 30 frames per second).

We perform k-fold cross-validation on every subject, with k chosen such that we always leave out 2 or 3 sequences for evaluation. All other sequences are used for training. We report the semantic IoU and image warping metric as introduced in the previous section, as well as landmarks reprojection errors averaged across subjects in Table 1. We also report per-subject results in our supplemental material. We outperform the competition numerically with similar inference times, which confirms our superior qualitative results.

4.3 Ablation Study

Number of sequences. To our knowledge, we present the first method that uses multiple in-the-wild monocular videos simultaneously for avatar reconstruction. To justify this choice, we ablate the number of sequences we use, both for training MultiFLARE and for the transfer learning. For 3 subjects for which we have at least 10 sequences, we leave 2 sequences out for evaluation and use 1, 2, 4 or 8 sequences for training. In Fig. 6 we show the reconstructed

Ablation	Warp PSNR ↑	Semantic IoU ↑	Landmarks L1 ↓
w/o transfer learning	30.073	0.621	0.042
w/o personalized appearance	30.449	0.685	0.028
Train MLP only	30.559	0.699	0.028
Train all	30.354	0.691	0.031
Ours	30.647	0.702	0.027

Table 2. Quantitative results of our ablation study, averaged over 6 subjects with cross-validation.

Sequences	Warp PSNR ↑	Semantic IoU ↑	Landmarks L1 ↓
1	31.574	0.665	0.037
2	31.659	0.681	0.036
4	31.911	0.695	0.033
8	32.028	0.706	0.030

Table 3. Ablation study on the numbers of sequences used for training. Results are averaged for 3 subjects.

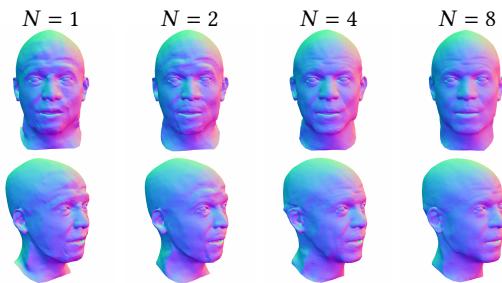


Fig. 6. Examples of reconstructed canonical geometry from MultiFLARE for an increasing number of training sequences.

neutral mesh obtained from MultiFLARE. In Table 3 we report the averaged IoU, image warping PSNR and landmarks reprojection errors over 3 subjects. Using more sequences increases our accuracy on all metrics, although we observe diminishing returns beyond 4. We attribute this improved performance to the more accurate canonical mesh, deformation model and material properties, and to the more diverse training set for transfer learning.

Positional encoding. We ablate the number of frequencies used by the positional encoding of the deformer \mathcal{D} (Eq. 2). In Fig. 7, we show the varying level of detail that we are able to reconstruct. Without positional encoding, the deformation network is unable to reconstruct the dynamic fine details of the geometry: these details are either lost or baked into the static canonical mesh. With additional frequencies, the fine details are much more defined and dynamic; however, too many high frequencies can lead to a noisy reconstruction.

Transfer learning. : In Table 2, we report quantitative results for different transfer learning options: training the last backbone layers and the MLPs (our proposed method), training the full backbones

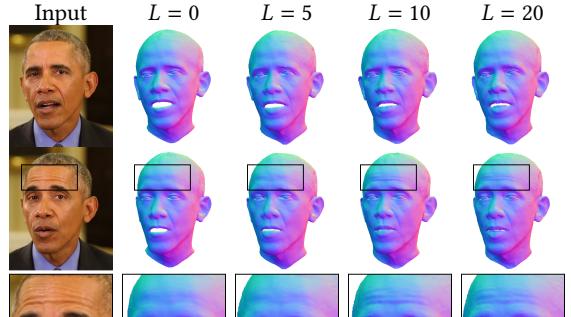


Fig. 7. Examples of reconstructed geometry using various levels of positional encoding in the deformer network (Eq 2) of MultiFLARE.

and MLPs, and training only the MLP. We find that freezing most of the backbone yields the best results: we keep the pre-trained backbone’s ability to generalize to diverse inputs while still allowing the highest-level features to adapt. We also evaluate the impact of using the appearance model estimated in MultiFLARE instead of the generic model from EMOCA. We show that our more precise appearance learned on the training sequences improves the performance of our final model.

5 LIMITATIONS AND FUTURE WORK

The personalized 3D geometry and appearance reconstruction stage of our method inherits several limitations that are specific to mesh-based avatar reconstruction methods. Occlusions and face accessories such as glasses cannot be represented accurately by our method. Furthermore, while we qualitatively show that minor facial hair can be handled, our canonical geometry and deformation model do not differentiate hair from skin, hence large beards and long hair cannot be accurately reconstructed. Similarly to FLARE [Bharadwaj et al. 2023], the simplified integration of the rendering equation, namely the split-sum approximation and pre-filtering of the environment map, results in difficulties for modeling sharp specular highlights and harsh shadows. In turn, this limits the accuracy of our personalized appearance model, thus harming the performance of our transfer-learning in such conditions.

The main limitations of the pre-trained face reconstruction encoder also apply to our method. While we do address some of the limitations that stem from the use of a parametric 3DMM, we are still subject to misalignment on extreme poses and strongly occluded faces.

Future work could explore further refinement of the personalization scheme and the modelling of per-sequence appearance variations to enhance the robustness and applicability of our method across a broader range of scenarios. Our work could also be applied to future state of the art face reconstruction methods, providing an even better starting point for our transfer learning stage.

6 ETHICS

This research is conducted with a focus on advancing the accuracy and realism of 3D face capture technologies for legitimate and ethical applications, such as improving visual effects in the entertainment

industry and enhancing user experiences in virtual and augmented reality. Our work is not intended for the creation of deepfakes or any unconsented face modifications. We firmly oppose the use of our method for malicious purposes, including the generation of misleading or harmful content. Our research is aimed solely at contributing to the scientific community and legitimate industry practices. We advocate for continued discussion and regulation to ensure that advancements in this field are used in ways that respect individual rights and community standards.

7 CONCLUSION

In this paper we present a methodology that performs a reconstruction of 3D face geometry and appearance from a collection of sources, which is then exploited to inform a transfer learning process. By replacing the decoder of a pre-trained monocular face reconstruction method with our personalized model, we achieve a more accurate self-supervision objective. This enables more precise expression and pose alignment, resulting in a new encoder capable of reconstructing 3D faces with expressive details from unseen images in real-time.

While current state-of-the-art methods offer real-time regression of parametric 3D face models, they fall short in providing the high-fidelity geometry required for high-end visual effects. By incorporating a more personalized approach, we address these limitations and demonstrate that leveraging unconstrained reference imagery can significantly enhance facial tracking quality.

Our results hold promising implications for the field of visual effects and other applications where reference material of a person is available. Overall, we contribute a novel and effective strategy for improving the fidelity of monocular face capture, paving the way for fully automated production-ready face capture.

8 ACKNOWLEDGMENTS

This project has received funding from the Association Nationale de la Recherche et de la Technologie under CIFRE agreement No 2022_1630.

REFERENCES

- Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. 2020. Cross-modal deep face normals with deactivable skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4979–4989.
- Victoria Fernández Abrevaya, Adnane Boukhayma, Stefanie Wuhrer, and Edmond Boyer. 2019. A decoupled 3d facial shape model by adversarial training. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9419–9428.
- Oswald Aldrian and William AP Smith. 2012. Inverse rendering of faces with a 3D morphable model. *IEEE transactions on pattern analysis and machine intelligence* 35, 5 (2012), 1080–1093.
- Riza Alp Guler, George Trigeorgis, Epameinondas Antonakos, Patrick Snape, Stefanos Zafeiriou, and Iasonas Kokkinos. 2017. Densereg: Fully convolutional dense shape regression in-the-wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6799–6808.
- Curtis Andrus, Junghyun Ahn, Michele Alessi, Abdallah Dib, Philippe Gosselin, Cédric Thébault, Louis Chevallier, and Marco Romeo. 2020. FaceLab: Scalable Facial Performance Capture for Visual Effects. In *The Digital Production Symposium (DigiPro '20)*. Association for Computing Machinery, New York, NY, USA, 1–3. <https://doi.org/10.1145/3403736.3403938>
- Anil Bas, William A. P. Smith, Timo Bolkart, and Stefanie Wuhrer. 2017. Fitting a 3D Morphable Model to Edges: A Comparison Between Hard and Soft Correspondences. In *Asian Conference on Computer Vision Workshops*. 377–391.
- Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. 2010. High-Quality Single-Shot Capture of Facial Geometry. *ACM Transactions on Graphics* 29, 4 (July 2010), 40:1–40:9. <https://doi.org/10.1145/1778765.1778777>
- Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. 2011. High-Quality Passive Facial Performance Capture using Anchor Frames. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*. 1–10.
- Joel Bennett and Chris Carter. 2014. Adopting virtual production for animated filmmaking. In *Proceedings of the 7th Annual International Conference on Computer Games, Multimedia and Allied Technology*. Global Science and Technology Forum (GSTF), 81–86.
- Shrish Bharadwaj, Yufeng Zheng, Otmar Hilliges, Michael J. Black, and Victoria Fernandez Abrevaya. 2023. FLARE: Fast learning of Animatable and Relightable Mesh Avatars. *ACM Transactions on Graphics* 42 (Dec. 2023), 15. <https://doi.org/10.1145/3618401>
- Volker Blanz, Sami Romdhani, and Thomas Vetter. 2002. Face identification across different poses and illuminations with a 3D morphable model. In *International Conference on Automatic Face & Gesture Recognition (FG)*. 202–207.
- Volker Blanz and Thomas Vetter. 1999. A Morphable Model for the Synthesis of 3D Faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*.
- Mario Botsch and Leif P. Kobbelt. 2004. A Remeshing Approach to Multiresolution Modeling. In *Eurographics Symposium on Geometry Processing*.
- Caio José Dos Santos Brito and Kenny Mitchell. 2019. Recycling a landmark dataset for real-time facial capture and animation with low cost HMD integrated cameras. In *Proceedings of the 17th International Conference on Virtual-Reality Continuum and its Applications in Industry*. 1–10.
- Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. 2015. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–9.
- Chen Cao, Qiming Hou, and Kun Zhou. 2014. Displaced dynamic expression regression for real-time facial tracking and animation. *Transactions on Graphics (TOG)* 33, 4 (2014), 1–10.
- Chen Cao, Yanlin Weng, Shun Zhou, Yiyi Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2013), 413–425.
- Feng-Ju Chang, Anh Tuan Tran, Tal Hassner, Iacopo Masi, Ram Nevatia, and Gerard Medioni. 2018. ExpNet: Landmark-free, deep, 3d facial expressions. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 122–129.
- Robert L Cook and Kenneth E. Torrance. 1982. A Reflectance Model for Computer Graphics. *ACM Transactions on Graphics (TOG)* 1, 1 (1982), 7–24.
- Radek Danecák, Michael Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 20279–20290. <https://doi.org/10.1109/CVPR52688.2022.01967>
- Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the Reflectance Field of a Human Face. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*. 145–156.
- Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5203–5212.
- Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. 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)*. 285–295.
- Pengfei Dou, Shishir K. Shah, and Ioannis A. Kakadiaris. 2017. End-to-end 3D face reconstruction with deep neural networks. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 5908–5917.
- Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from-in-the-wild images. *Transactions on Graphics (Proc. SIGGRAPH)* 40, 4 (2021), 1–13.
- Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. 2018. Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression Network. In *European Conference on Computer Vision (ECCV)*.
- Panagiotis P. Filntisis, George Retsinas, Foivos Paraperas-Papantonio, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. 2023. SPECTRE: Visual Speech-Informed Perceptual 3D Facial Expression Reconstruction From Videos. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*. 5745–5755.
- Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8649–8658.
- Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* 132, 2 (2024), 581–595.

- Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. 2022. Reconstructing Personalized Semantic Facial NeRF Models From Monocular Video. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 41, 6 (2022).
- Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. 2016. Reconstruction of personalized 3D face rigs from monocular video. *ACM Transactions on Graphics (TOG)* 35, 3 (2016), 1–15.
- Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. 2018. Unsupervised Training for 3D Morphable Model Regression. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 8377–8386.
- Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schoenborn, and Thomas Vetter. 2018. Morphable Face Models - An Open Framework. In *International Conference on Automatic Face & Gesture Recognition (FG)*, 75–82. <https://doi.org/10.1109/FG.2018.00021>
- Abhijeet Ghosh, Graham Kyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview Face Capture using Polarized Spherical Gradient Illumination. *ACM Transactions on Graphics (TOG)* 30, 6 (2011), 1–10.
- Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural Head Avatars from Monocular RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. 2022. Shape, light, and material decomposition from images using monte carlo rendering and denoising. *Advances in Neural Information Processing Systems* 35 (2022), 22856–22869.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Pablo Helman, Leandro Estebecorena, and Stephane Grabli. 2020. ILM presents: making "The Irishman". In *ACM SIGGRAPH 2020 Production Sessions*. 1–1.
- Aaron S Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE international conference on computer vision*, 1031–1039.
- Harim Jung, Myeong-Seok Oh, and Seong-Whan Lee. 2021. Learning Free-Form Deformation for 3D Face Reconstruction from In-The-Wild Images. In *International Conference on Systems, Man, and Cybernetics (SMC)*. 2737–2742. <https://doi.org/10.1109/SMC52423.2021.9659124>
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics* 42, 4 (2023), 1–14.
- Hyeyoungwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. 2018. InverseFaceNet: Deep Monocular Inverse Face Rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4625–4634.
- Tatsuro Koizumi and William A. P. Smith. 2020. "Look Ma, No Landmarks!" - Unsupervised, Model-Based Dense Face Alignment. In *European Conference on Computer Vision (ECCV)*. Vol. 12347. 690–706.
- Samuli Laine, Tero Karras, Timo Aila, Antti Herva, Shunsuke Saito, Ronald Yu, Hao Li, and Jaakko Lehtinen. 2017. Production-level facial performance capture using deep convolutional neural networks. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*. 1–10.
- Yoonho Lee, Annie S Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn. 2022. Surgical fine-tuning improves adaptation to distribution shifts. *arXiv preprint arXiv:2210.11466* (2022).
- Hao Li, Jihun Yu, Yuting Ye, and Chris Bregler. 2013. Realtime facial animation with on-the-fly correctives. *Transactions on Graphics (TOG)* 32, 4 (2013), 42–1.
- Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. <https://doi.org/10.1145/3130800.3130813>
- Zhaoshou Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yaojie Liu, Amin Jourabloo, William Ren, and Xiaoming Liu. 2017. Dense Face Alignment. In *International Conference on Computer Vision Workshops (ICCV-W)*. 1619–1628.
- Shengjie Ma, Yanlin Weng, Tianjia Shao, and Kun Zhou. 2024. 3D Gaussian Blendshapes for Head Avatar Animation. *arXiv preprint arXiv:2404.19398* (2024).
- Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 41, 4, Article 102 (July 2022), 15 pages.
- Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. 2022. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8280–8290.
- Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *2009 sixth IEEE international conference on advanced video and signal based surveillance*. Ieee, 296–301.
- Stylianos Ploumpis, Evangelos Ververas, Eimear O' Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick E. Pears, William A. P. Smith, Baris Gececi, and Stefanos Zafeiriou. 2021. Towards a Complete 3D Morphable Model of the Human Head. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 43, 11 (2021), 4142–4160.
- Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Gibenbain, and Matthias Nießner. 2023. Gaussian Avatars: Photorealistic Head Avatars with Rigged 3D Gaussians. *arXiv preprint arXiv:2312.02069* (2023).
- George Retsinas, Panagiotis P. Filntisis, Radek Danecek, Victoria F. Abrevaya, Anastasios Roussos, Timo Bolkart, and Petros Maragos. 2024. SMIRK: 3D Facial Expressions through Analysis-by-Neural-Synthesis. *arXiv:2404.04104* (April 2024). [arXiv:2404.04104](https://arxiv.org/abs/2404.04104)
- E. Richardson, M. Sela, and R. Kimmel. 2016. 3D Face Reconstruction by Learning from Synthetic Data. In *International Conference on 3D Vision (3DV)*. 460–469.
- Jérémie Rivière, Pablo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-Shot High-Quality Facial Geometry and Skin Appearance Capture. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*. Association for Computing Machinery (ACM).
- Sami Romdhani and Thomas Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and apriori. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, Vol. 2. 986–993.
- Zeyu Ruan, Changqing Zou, Longhai Wu, Gangshan Wu, and Limin Wang. 2021. SADRNet: Self-aligned dual face regression networks for robust 3D dense face alignment and reconstruction. *IEEE Transactions on Image Processing* 30 (2021), 5793–5806.
- Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. 2019. Learning to Regress 3D Face Shape and Expression from an Image without 3D Supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 7763–7772.
- Matan Sela, Elad Richardson, and Ron Kimmel. 2017. Unrestricted facial geometry reconstruction using image-to-image translation. In *International Conference on Computer Vision (ICCV)*. 1576–1585.
- Jiaxiang Shang, Tianwei Shen, Shiwei Li, Lei Zhou, Mingmin Zhen, Tian Fang, and Long Quan. 2020. Self-Supervised Monocular 3D Face Reconstruction by Occlusion-Aware Multi-view Geometry Consistency. In *European Conference on Computer Vision (ECCV)*, Vol. 12360. Springer, 53–70.
- Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*.
- Attila Szabó, Givi Meishvili, and Paolo Favaro. 2019. Unsupervised Generative 3D Shape Learning from Natural Images. *CoRR abs/1910.00287* (2019). <http://arxiv.org/abs/1910.00287>
- Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgarib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. 2019. FML: Face Model Learning from Videos. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 10812–10822.
- Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2549–2559.
- Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. 2017. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *International Conference on Computer Vision (ICCV)*. 1274–1283.
- Justus Thies, Michael Zollhöfer, Matthias Nießner, Levi Valgaerts, Marc Stamminger, and Christian Theobalt. 2015. Real-Time Expression Transfer for Facial Reenactment. *ACM Trans. Graph.* 34, 6 (2015).
- Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016a. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.
- Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016b. Facevr: Real-time facial reenactment and eye gaze control in virtual reality. *arXiv preprint arXiv:1610.03151* (2016).

- Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gerard Medioni. 2017. Regressing Robust and Discriminative 3D Morphable Models With a Very Deep Neural Network. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 1599–1608.
- Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. 2018. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3935–3944.
- Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. 2022. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. *CVPR* (2022).
- Huawei Wei, Shuang Liang, and Yichen Wei. 2019. 3D Dense Face Alignment via Graph Convolution Networks. *arXiv preprint arXiv:1904.05562* (2019).
- Erroll Wood, Tadas Baltrusaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljevic, Daniel Wilde, Stephan J. Garbin, Toby Sharp, Ivan Stojiljkovic, Tom Cashman, and Julien P. C. Valentin. 2022. 3D Face Reconstruction with Dense Landmarks. In *European Conference on Computer Vision (ECCV)*, Vol. 13673. Springer, 160–177.
- Chenglei Wu, Takaaki Shiratori, and Yasir Sheikh. 2018. Deep incremental learning for efficient high-fidelity face tracking. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–12.
- Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. 2020. Unsupervised Learning of Probably Symmetric Deformable 3D Objects From Images in the Wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 1–10.
- Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. 2024. Gaussian Head Avatar: Ultra High-fidelity Head Avatar via Dynamic Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Yuelang Xu, Lizhen Wang, Xiaochen Zhao, Hongwen Zhang, and Yebin Liu. 2023. AvatarMAV: Fast 3D Head Avatar Reconstruction Using Motion-Aware Neural Voxels. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*.
- Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. 2021. i3dmm: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12803–12813.
- Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. 2018. BiSeNet: Bilateral Segmentation Network for Real-Time Semantic Segmentation. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.), Vol. 11217. Springer International Publishing, Cham, 334–349. https://doi.org/10.1007/978-3-030-01261-8_20
- Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. 2019. DF2Net: A Dense-Fine-Finer Network for Detailed 3D Face Reconstruction. In *International Conference on Computer Vision (ICCV)*.
- Tianke Zhang, Xiangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, et al. 2023. Accurate 3D Face Reconstruction with Facial Component Tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9033–9042.
- Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. 2022. I M Avatar: Implicit Morphable Head Avatars from Videos. In *Computer Vision and Pattern Recognition (CVPR)*.
- Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, and Otmar Hilliges. 2023. PointAvatar: Deformable Point-based Head Avatars from Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.
- Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. 2016. Face Alignment Across Large Poses: A 3D Solution. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 146–155.
- Wojciech Zienonka, Timo Bolkart, and Justus Thies. 2022. Towards Metrical Reconstruction of Human Faces. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIII*. Springer-Verlag, Berlin, Heidelberg, 250–269. https://doi.org/10.1007/978-3-031-19778-9_15
- Wojciech Zienonka, Timo Bolkart, and Justus Thies. 2023. Instant Volumetric Head Avatars. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Received 19 May 2024; revised 6 September 2024; accepted ?

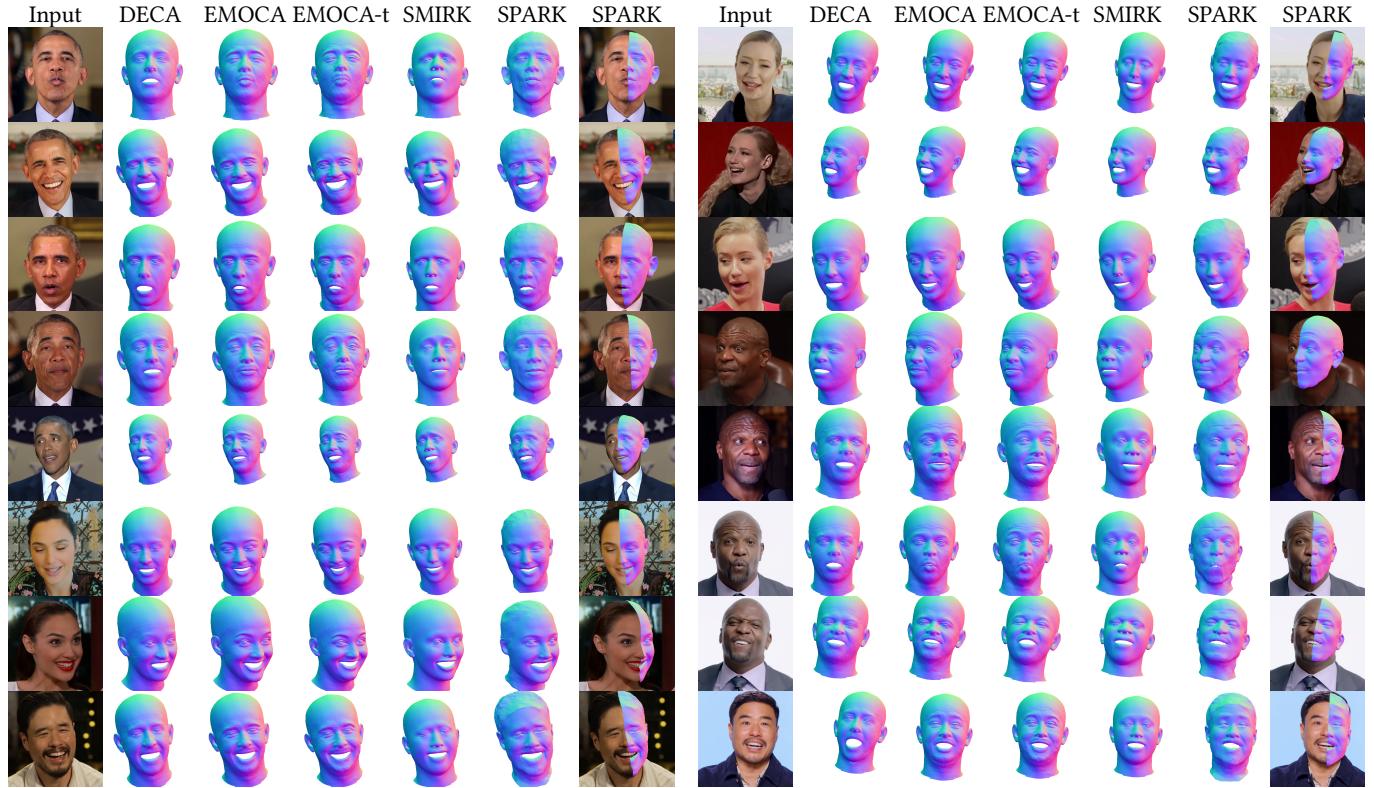


Fig. 8. Additional results of our method compared to multiple face reconstruction methods. From left to right: input image, DECA [Feng et al. 2021], EMOCA [Danecek et al. 2022], EMOCA fine-tuned, SMIRK [Retsinas et al. 2024] and SPARK.

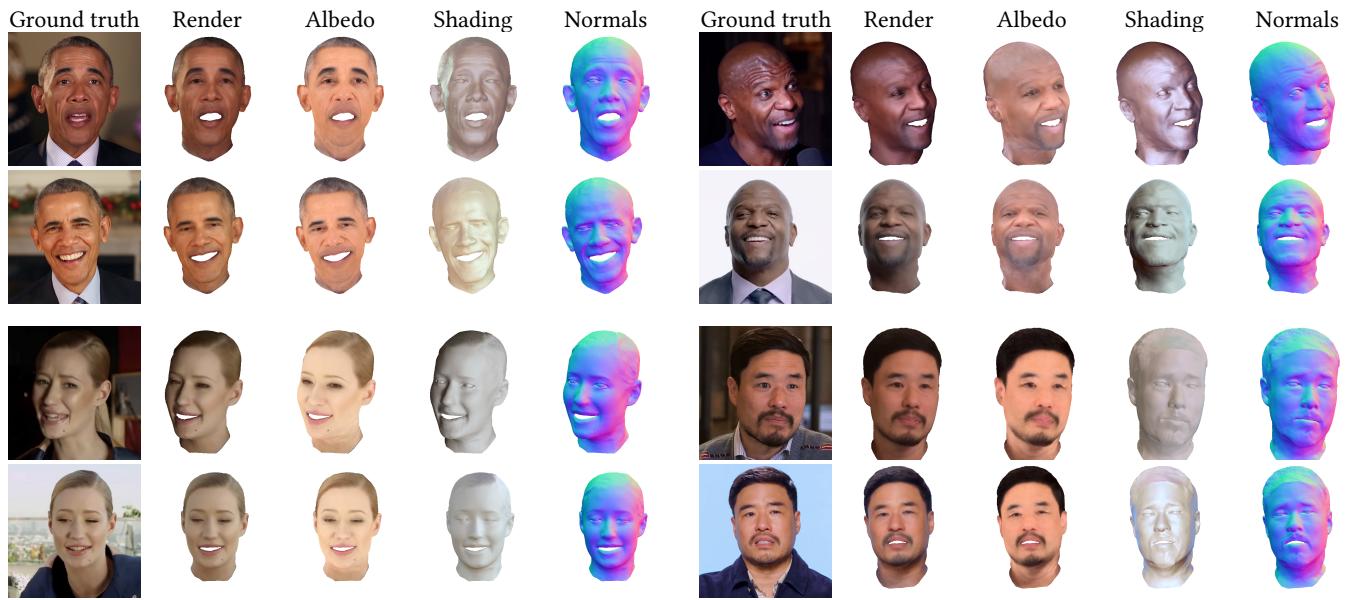


Fig. 9. Qualitative results from MultiFLARE. For a given subject, only the shading varies between sequences. While we only show samples for two sequences per subject, the method was trained with 4 to 6 sequences for each subject.

SPARK: Self-supervised Personalized Real-time Monocular Face Capture

Supplementary Material

KELIAN BAERT, Technicolor Group, France and University Rennes, France
SHRISHA BHARADWAJ, Max Planck Institute for Intelligent Systems, Germany
FABIEN CASTAN, Technicolor Group, France
BENOIT MAUJEAN, Technicolor Group, France
MARC CHRISTIE, University Rennes, IRISA, CNRS, Inria, France
VICTORIA F. ABREVAYA, Max Planck Institute for Intelligent Systems, Germany
ADNANE BOUKHAYMA, Inria, University Rennes, IRISA, CNRS, France

ACM Reference Format:

Kelian Baert, Shrisha Bharadwaj, Fabien Castan, Benoit Maujean, Marc Christie, Victoria F. Abrevaya, and Adnane Boukhayma. 2024. SPARK: Self-supervised Personalized Real-time Monocular Face Capture **Supplementary Material**. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24)*, December 3–6, 2024, Tokyo, Japan. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3680528.3687704>

1 INTRODUCTION

In this supplemental document, we provide additional details on our training process, sources for our dataset and detailed quantitative results for each subject.

2 DATASET

We compile our dataset from online videos, for which we provide sources in Table 1. We manually cut the videos to only include a single viewpoint of the subject, removing frames where the person is heavily occluded or not visible at all. In several cases, we sample multiple sequences from the same source video. For the purposes of our work, we treat those as different videos because of the entirely different illumination conditions.

3 TRAINING DETAILS

In this section, we provide more information on the training of both stages of our method.

3.1 MultiFLARE

We train MultiFLARE for 3000 iterations using a batch size of 4 images. The canonical geometry is remeshed [Botsch and Kobbelt 2004] to increase its resolution after 500 iterations. We use an Adam optimizer [Kingma and Ba 2015] with a learning rate of 2e-4 for the deformation network \mathcal{D} and canonical vertices x_c , 1e - 3 for the material network \mathcal{M} and light MLPs \mathcal{L}_i . We also optimize pose and expression parameters with a learning rate of 1e-4. Training takes about 20 minutes on an NVIDIA RTX A5000 GPU.

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24)*, December 3–6, 2024, Tokyo, Japan, <https://doi.org/10.1145/3680528.3687704>.

Deformation network. The deformation network \mathcal{D} is initialized through a quick supervised learning, minimizing $\|\mathcal{D}(\gamma(x)) - \mathcal{E}\|_2$ at the positions of the FLAME [Li et al. 2017] template vertices. The deformer will initially act as a continuous representation of the FLAME expression basis \mathcal{E} independent of topology. Note that we only observed a marginal difference when subdividing the FLAME template mesh for this stage, as the MLP's inductive bias already provides the deformer interpolation ability. We use the Adam optimizer with a learning rate of 2e-4 and train for 5000 iterations (using all vertex positions of the template mesh at every iteration), requiring under a minute on a NVIDIA RTX A5000 GPU.

Progressive multi-resolution hash encoding. We embed the input of the material network \mathcal{M} using multi-resolution hash encoding [Müller et al. 2022], with 16 resolutions levels ranging from 16 to 4096 and 2 features per level. These increasingly detailed levels are progressively enabled [Li et al. 2023]: one every 250 training iterations, and the remaining 8 at iteration 2000.

Network architectures. The material network \mathcal{M} and the illumination networks \mathcal{L}_i all have 3 hidden layers of size 64, with ReLU activations. The final prediction of \mathcal{M} is activated using Softplus ($\beta = 100$), while \mathcal{L}_i outputs use a Sigmoid activation. The deformer \mathcal{D} has 4 hidden layers of size 128 with Softplus activations ($\beta = 100$).

3.2 Tracker adaptation

We perform the transfer-learning stage using the EMOCA v2 [Danecek et al. 2022; Filntsis et al. 2023] encoder. We freeze most of the encoder and only update the weights of the last ResNet-50 [He et al. 2016] block and the MLP head, for both the coarse shape encoder and the expression encoder. We train for a maximum of 30 epochs using the Adam optimizer with a learning rate of 1e - 5 and a batch size of 16, using early stopping to minimize the risk of overfitting. We set weights λ_{emo} , λ_{eye} , λ_{mc} , λ_ψ and λ_{lip} to the values provided by the authors' implementation and increase λ_{pho} from 2 to 100. Doing so, we take advantage of our more accurate pre-estimated appearance model from MultiFLARE. On our datasets, training is completed in 20 minutes on average, depending on early stopping.

The cumulative training time for both the MultiFLARE and the transfer-learning stage is less than 45 minutes. Once trained, the expression basis from the deformation network can be precomputed and the convolutional encoder, a ResNet-50 with a MLP head, can perform inference in real-time on modern hardware.

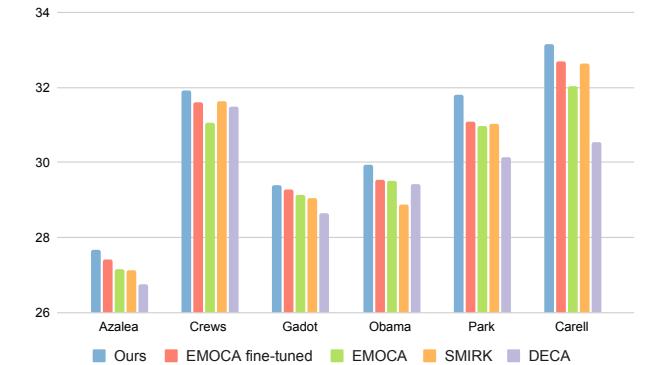
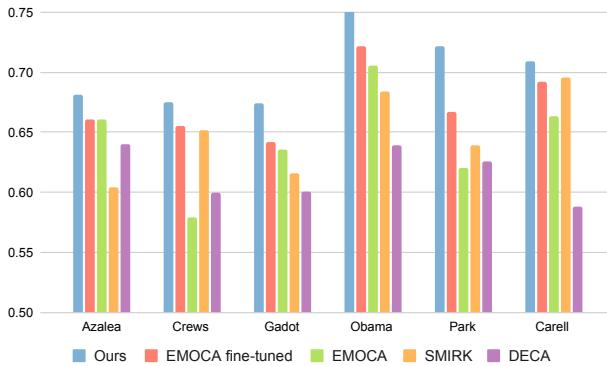


Fig. 1. Results on 6 datasets for our semantic IoU (top) and geometry-based image warping PSNR (bottom) metrics, compared against several state-of-the-art methods. For each dataset, we perform cross-validation and average the results.

4 DETAILED RESULTS

We report detailed per-subject results of our final personalized method for the semantic IoU and image warping metrics. Fig. 1 shows a comparison of our method against state-of-the-art approaches for monocular face reconstruction. Fig. 2 shows an ablation study of our method for various training schemes in the tracker adaptation stage.

REFERENCES

- Mario Botsch and Leif P. Kobbelt. 2004. A Remeshing Approach to Multiresolution Modeling. In *Eurographics Symposium on Geometry Processing*.
- Radek Danecek, Michael Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 20279–20290. <https://doi.org/10.1109/CVPR52688.2022.01967>
- Panagiotis P. Filntasis, George Retsinas, Foivos Parperas-Papantoniu, Athanasios Katsamanis, Anastasios Roussos, and Petros Maragos. 2023. SPECTRE: Visual Speech-Informed Perceptual 3D Facial Expression Reconstruction From Videos. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPR-W)*. 5745–5755.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Las Vegas, NV, USA, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Diederik P Kingma and Jimmy Ba. 2015. ADAM: A Method for Stochastic Optimization. In *International Conference on Learning Representations (ICLR)*.

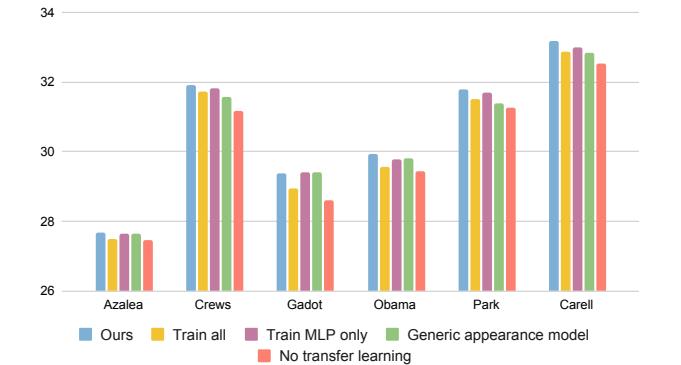
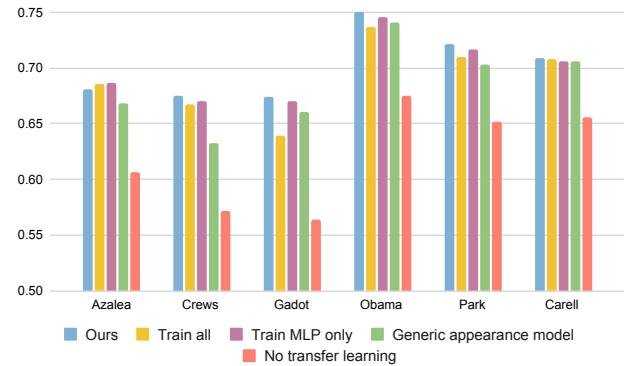


Fig. 2. Results of our ablation study on 6 datasets for our semantic IoU (top) and geometry-based image warping PSNR metrics (bottom). For each dataset, we perform cross-validation and average the results.

- Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. <https://doi.org/10.1145/3130800.3130813>
- Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. 2023. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 41, 4, Article 102 (July 2022), 15 pages.

Subject	Video	Cut duration	Source
Barack Obama	1	01:00	https://www.youtube.com/watch?v=7G5kMmnAp_8
	2	00:43	https://www.youtube.com/watch?v=1CW6NbZvR_w
	3	01:00	https://www.youtube.com/watch?v=lbwlVwNWLzU
	4	01:00	https://www.youtube.com/watch?v=PtkBjuQblqI
	5	01:00	https://www.youtube.com/watch?v=-XecjTOrNs
	6	00:56	https://www.youtube.com/watch?v=_iKNE4ndBdg
	7	01:00	https://www.youtube.com/watch?v=siyBp8Csugk
	8	01:00	https://www.youtube.com/watch?v=aip0BAWrldLw
	9	01:00	https://www.youtube.com/watch?v=1Z9tiSqHzkg
	10	01:00	https://www.youtube.com/watch?v=BloBwbllxws
Iggy Azalea	1	00:29	
	2	00:40	
	3	00:27	https://www.youtube.com/watch?v=AMh5f8xRLRE
	4	00:44	
	5	00:43	https://www.youtube.com/watch?v=-jlPXv6Ehks
	6	00:30	https://www.youtube.com/watch?v=YvB1o9wCO80
	7	00:40	https://www.youtube.com/watch?v=AogzdSQ5LVE
Gal Gadot	1	00:32	
	2	00:13	
	3	00:29	https://www.youtube.com/watch?v=OJJMVLPdAwY
	4	00:30	
	5	00:17	https://www.youtube.com/watch?v=TJVf3KZAIG4
	6	00:12	https://www.youtube.com/watch?v=iX01L8wmhBk
Terry Crews	1	01:12	https://www.youtube.com/watch?v=eM1XfAsGnHI
	2	00:42	https://www.youtube.com/watch?v=BvA_AinclK8
	3	00:47	https://www.youtube.com/watch?v=fJjsbgk6e0A
	4	00:33	https://www.youtube.com/watch?v=Im8itGAyZ6M
	5	00:59	https://www.youtube.com/watch?v=hJeg601TIU4
	6	01:07	https://www.youtube.com/watch?v=iuX7NZ56Ei4
	7	00:49	https://www.youtube.com/watch?v=PnoCKY7L-s
	8	01:06	https://www.youtube.com/watch?v=v0F3BhonHMM
	9	00:28	https://www.youtube.com/watch?v=PmbXC3dlK_s
	10	00:58	https://www.youtube.com/watch?v=ITzb_I0y_EM
Randall Park	1	00:32	https://www.youtube.com/watch?v=LKqvgRAZjZg
	2	00:59	https://www.youtube.com/watch?v=nbfseMeA9n4
	3	01:13	https://www.youtube.com/watch?v=wepD5rUZf8k
	4	01:03	https://www.youtube.com/watch?v=qT7nP4zQ05Y
	5	00:43	https://www.youtube.com/watch?v=mTPchtLBjJI
	6	00:59	https://www.youtube.com/watch?v=nbY34h54av0
	7	00:50	https://www.youtube.com/watch?v=2eQN4nCW25c
	8	00:58	https://www.youtube.com/watch?v=2eQN4nCW25c
Steve Carell	1	00:23	
	2	00:16	
	3	00:22	
	4	00:10	
	5	00:15	
	6	00:09	<i>The Office</i> season 2
	7	00:08	
	8	00:27	
	9	00:18	
	10	00:30	
	11	00:19	
	12	00:29	

Table 1. Sources for all videos of our dataset.