Decaf: Monocular Deformation Capture for Face and Hand Interactions

SOSHI SHIMADA, MPI for Informatics, SIC, and VIA Research Center, Germany
VLADISLAV GOLYANIK, MPI for Informatics and SIC, Germany
PATRICK PÉREZ, Valeo.ai, France
CHRISTIAN THEOBALT, MPI for Informatics, SIC, and VIA Research Center, Germany

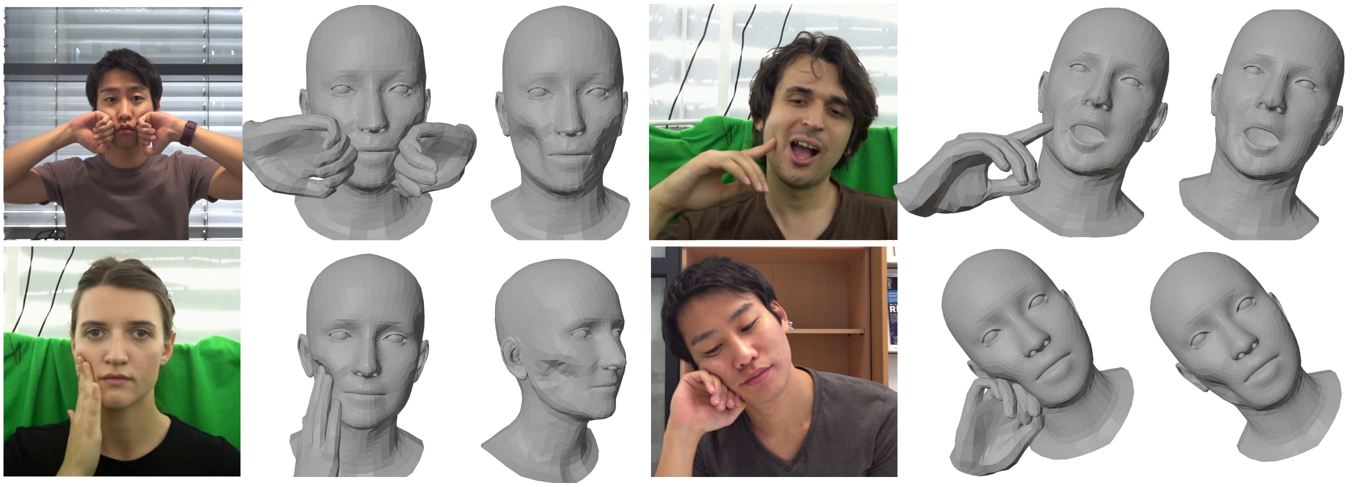

Fig. 1. **Our *Decaf* approach captures hands and face motions as well as the *face surface deformations* arising from the interactions from a single-view RGB video.** Thanks to our new dataset with 3D surface deformations relying on position based dynamics that considers the underlying human skull structure, our neural architecture estimates plausible hands-head interactions and head deformations. The examples in this figure highlight the variety of the supported hand poses and facial expressions. The results are temporally consistent. See our supplementary video for dynamic visualisations.

Existing methods for 3D tracking from monocular RGB videos predominantly consider articulated and rigid objects (*e.g.,* two hands or humans interacting with rigid environments). Modelling dense non-rigid object deformations in this setting (*e.g.* when hands are interacting with a face), remained largely unaddressed so far, although such effects can improve the realism of the downstream applications such as AR/VR, 3D virtual avatar communications, and character animations. This is due to the severe ill-posedness of the monocular view setting and the associated challenges (*e.g.,* in acquiring a dataset for training and evaluation or obtaining the reasonable non-uniform stiffness of the deformable object). While it is possible to naïvely track multiple non-rigid objects independently using 3D templates or parametric 3D models, such an approach would suffer from multiple artefacts in the resulting 3D estimates such as depth ambiguity, unnatural intra-object collisions and missing or implausible deformations.

Hence, this paper introduces the first method that addresses the fundamental challenges depicted above and that allows tracking human hands interacting with human faces in 3D from single monocular RGB videos. We model hands as articulated objects inducing non-rigid face deformations during an active interaction. Our method relies on a new hand-face motion and interaction capture dataset with realistic face deformations acquired with a markerless multi-view camera system. As a pivotal step in its creation, we process the reconstructed raw 3D shapes with position-based dynamics and an approach for non-uniform stiffness estimation of the head tissues, which results in plausible annotations of the surface deformations, hand-face contact regions and head-hand positions. At the core of our neural approach are a variational auto-encoder supplying the hand-face depth prior and modules that guide the 3D tracking by estimating the contacts and the deformations. Our final 3D hand and face reconstructions are realistic and more plausible compared to several baselines applicable in our setting, both quantitatively and qualitatively. https://vcai.mpi-inf.mpg.de/projects/Decaf

CCS Concepts: • **Computing methodologies** → **Computer graphics**; **Motion capture**.

Additional Key Words and Phrases: monocular, motion capture, interaction, deformation

ACM Reference Format:
Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. 2023. Decaf: Monocular Deformation Capture for Face and Hand Interactions. *ACM Trans. Graph.* 42, 6, Article 264 (December 2023), 16 pages. https://doi.org/10.1145/3618329

Authors' addresses: Soshi Shimada, MPI for Informatics and SIC and VIA Research Center, Saarbrücken, Germany, sshimada@mpi-inf.mpg.de; Vladislav Golyanik, MPI for Informatics and SIC, Saarbrücken, Germany, golyanik@mpi-inf.mpg.de; Patrick Pérez, Valeo.ai, Paris, France, patrick.perez@valeo.com; Christian Theobalt, MPI for Informatics and SIC and VIA Research Center, Saarbrücken, Germany, theobalt@mpi-inf.mpg.de.

0730-0301/2023/12-ART264
https://doi.org/10.1145/3618329

1 INTRODUCTION

Reconstructing 3D hands and face from a **monocular RGB video** is a challenging and important research area in computer graphics. The task becomes significantly more difficult when attempting to reconstruct hands and face simultaneously including *surface deformations caused by their interactions*. Capturing such interactions and deformations is crucial for enhancing realism in reconstructions as they are frequently observed in everyday life (hand-face interaction occurs 23 times per hour on average during awake-time [Kwok et al. 2015]), and they significantly impact the impressions formed by others. Consequently, reconstructing hand-face interactions is key for avatar communication, virtual/augmented reality, and character animation, where realistic facial movements are essential to create an immersive experience, as well as for applications such as sign language transcriptions and driver drowsiness monitoring. Despite several studies on the reconstruction of face and hand motions, the capture of interactions between them and the corresponding deformations from a monocular RGB video remains unaddressed [Tretschk et al. 2023]. On the other hand, naïvely using existing template-based hand and face reconstruction methods leads to artefacts such as collisions, and missing interactions and deformations due to the inherent depth ambiguity in the monocular setting and the lack of deformation modelling in the reconstruction pipeline.

Several key challenges are associated with this problem setting. One (I) is the lack of an available markerless RGB capture dataset for face and hand interaction with non-rigid deformations for model training and method evaluation. Capturing such a dataset is highly challenging due to the constant presence of occlusions caused by hand and head motions, particularly at the interaction region where non-rigid deformation occurs. Another challenge (II) is the inherent depth ambiguity of the single-view RGB setup, which makes it difficult to obtain accurate localisation information, resulting in errors that can cause implausible artefacts such as collisions or non-touching of the hand and head (when they interact in practice). To tackle these challenges, we propose *Decaf* (short for *deformation capture of faces interacting with hands*), a monocular RGB method for capturing face and hand interactions along with facial deformations.

Specifically, to address (I), we propose a solution that combines a multiview capture setup with a position-based dynamics simulator for reconstructing the interacting surface geometry, even under occlusions. To integrate the deformable object simulator, we calculate the stiffness values of a head mesh using a simple but effective“skull-skin distance” (SSD) method. This approach provides non-uniform stiffness to the mesh, which significantly improves the qualitative plausibility of the reconstructed geometry compared with uniform stiffness values. To address the challenge (II), we train the networks to obtain the 3D surface deformations, contact regions on the head and hand surfaces, and the interaction depth prior from single-view RGB images utilising our new dataset. During the final optimisation stage, we utilise this information from different modalities to obtain plausible 3D hand and face interactions with non-rigid surface deformations, which helps disambiguate the depth ambiguity of the single-view setup. Our approach results in much more plausible hands-face interactions compared to the existing works; see Fig. 1 for representative results.

In summary, the primary technical contributions of this article are as follows:

- *Decaf*, the first learning-based MoCap approach for 3D hand and face interaction reconstruction with face surface deformations (Sec. 3).
- A global fitting optimisation guided by the estimated contacts, learned interaction depth prior, and deformation model of the face to enable plausible 3D interactions (Sec. 3.3).
- The acquisition of the first markerless RGB-based 3D hand-face interaction dataset with surface deformations with consistent topology based on position-based dynamics (PBD). The reference 3D data for model training and evaluation are generated using a simple and effective non-uniform stiffness estimation approach for human head models, namely *skull-skin distance* (*SSD*; Sec. 4).

Our *Decaf* outperforms benchmark and existing related methods both qualitatively and quantitatively, with notable improvements in physical plausibility metrics (Sec. 5.3). For dynamic qualitative comparisons, please refer to our supplementary video. We plan to release the acquired dataset and code for research purposes.

2 RELATED WORKS

This section focuses on the 3D reconstruction of hands interacting with objects in the monocular (single-view) capture context.

2.1 Hand Reconstruction with Interactions

There have been diverse works proposed to capture 3D hand motions with interactions. Several works reconstruct 3D hand and rigid object interactions from depth information [Hu et al. 2022; Zhang et al. 2019, 2021b] or RGB camera [Cao et al. 2021; Grady et al. 2021; Liu et al. 2021; Tekin et al. 2019]. There are several works that reconstruct hand-hand interactions. Mueller *et al.* [2019] reconstruct two hands interactions from a single depth camera utilising collision proxies based on Gaussian spheres embedded in the hand model. Some works reconstruct interacting 3D hands from a single RGB image [Wang et al. 2022; Zhang et al. 2021a]. However, none of these works considers the non-rigidity while interactions unlike ours.

Similar to our approach, Tsoli *et al.* [2018] reconstruct **non-rigid** cloth and hand interaction by considering hand/object contact points in the optimisation. However, the method requires **RGB-D** input unlike ours. Our work assumes no access to depth sensor information and reconstructs interactions with a deformable face. The face exhibits varying stiffness values based on the surface area, owing to the underlying skull structure in a human's head. This is in contrast to cloth interactions, which typically have uniform stiffness values. Furthermore, our face autonomously changes its pose and expression during the sequence, whereas in [Tsoli and Argyros 2018], the behaviour of the cloth changes only due to the interacting hand or gravity. These unique characteristics, coupled with the limited input setting, make our problem highly challenging.

2.2 Monocular Face Reconstruction

Capturing a human face from a single view RGB input is important for many graphics applications, thus a significant amount of works

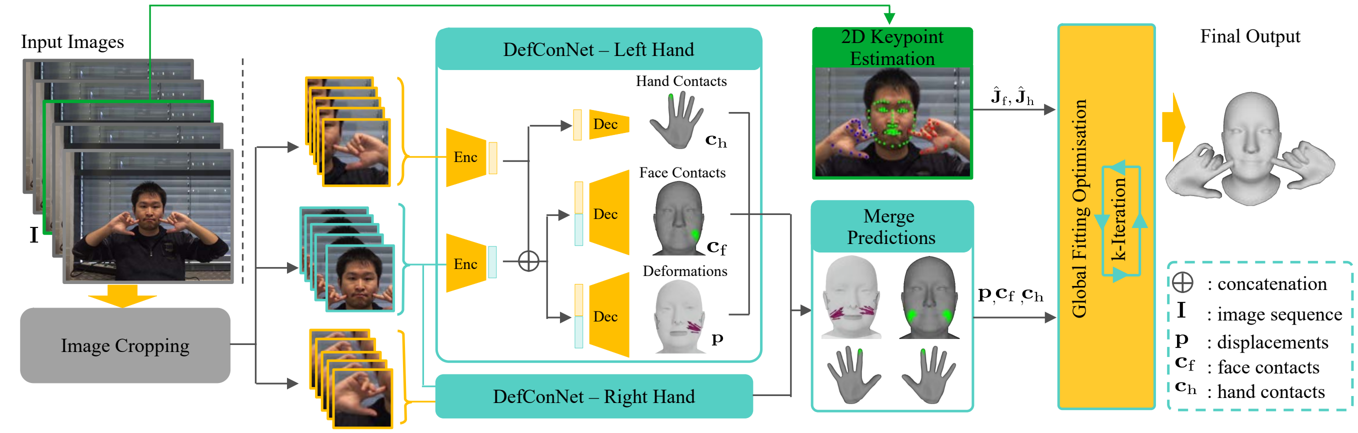

Fig. 2. Schematic visualisation of *Decaf*, the proposed system to predict 3D poses of hands and face in interaction from a sequence of monocular RGB images of a subject. The input image sequence is first cropped on the left-/right-hand and face locations, which are subsequently fed to the DefConNet, where we estimate the probabilities of face-hand contact $\mathbf{c}_\mathrm{f}$ and $\mathbf{c}_\mathrm{h}$ as well as the per-vertex displacements $\mathbf{p}_0$. Two DefConNets of the same architecture are independently operated in cases where two hands are present in the scene. Next, the contact labels and deformations are merged by simply computing the union of the outputs from the two DecConNets. Finally, we solve the optimisation to fit the hands and face parametric models guided by the estimated contacts, deformations and 2D keypoints in the image (Sec. 3.3). The final output from *Decaf* reconstructs the face and hands, incorporating plausible surface deformations on the face resulting from their interactions.

have been proposed with learning-free [Garrido et al. 2013, 2016; Thies et al. 2016; Wu et al. 2016] and learning-based approaches [Ichim et al. 2015; Lattas et al. 2020; Saito et al. 2016]. In this category, some works train the networks in a self-supervised manner to reconstruct faces with textures and illuminations [Tewari et al. 2017] or details with estimated normals [Danecek et al. 2022; Feng et al. 2021b]. Although these works capture the geometry of expressive deforming human faces, none of the works in this category models the face deformations caused by the interactions unlike ours.

2.3 Shape from Template (SfT)

This algorithm class bears a similarity to our approach. SfT assumes a template mesh of the tracking object and deforms the template mesh based on the observations such as RGB/-D sequences. Several works address this problem with learning-based algorithms [Bozic et al. 2020; Fuentes-Jimenez et al. 2021; Golyanik et al. 2018; Kairanda et al. 2022; Shimada et al. 2019], and some with learning-free optimisation-based approaches [Habermann et al. 2018; Ngo et al. 2015; Salzmann et al. 2007; Yu et al. 2015; Zollhöfer et al. 2014]. Unlike these approaches, our method models *interactions* between two different objects (*i.e.* hand and face) from a single view RGB input under severe occlusions caused by the interactions. Petit *et al.* [2018] propose a physics-based non-rigid object tracking method using a finite element method. However, their method requires RGB-D input and focuses on simple deformable objects (*e.g.* , cubes and discs). In contrast, our approach does not rely on depth information and handles interactions between a complex articulated hand and face, considering locally varying stiffness values. Some works estimate 3D human poses with self- and multi-person interactions (contacts) from single RGB images [Fieraru et al. 2020, 2021; Müller et al. 2021]. However, they do not model significant surface deformations due to contacts (*e.g.* during hand-face interactions). Li *et al.* [2022] propose a method that addresses a problem set that bears resemblance to ours. It estimates the 3D global human pose along with the deformations of the interacting environment surface based on ARAP-loss. However, their method does not consider stiffness values specific to object categories and does not incorporate learned priors for non-rigid deformations, distinguishing it from our approach.

2.4 Template Free Non-Rigid Surface Tracking

Some methods in this category reconstruct non-rigid surfaces by acquiring first an explicit template mesh from RGB-D inputs [Innmann et al. 2016]. Some use node graphs [Lin et al. 2022] or implicit SDF surface representations [Slavcheva et al. 2017] for non-rigid surface tracking. Guo *et al.* [2017] propose a method that reconstructs the non-rigid surface along with the surface albedo and low-frequency lighting. Our approach differs from these works by considering the dynamics of the interactions between two different materials *i.e.* face and hand, and face surface stiffness values based on bone structure. Additionally, our dataset and method's output have consistent 3D mesh topologies that are very important for the supervision of network training in explicit surface space.

2.5 Physics-based MoCap

Recently, numerous physics-based algorithms for motion capture have been proposed. Several works model the interactions with the environment from a static single RGB camera [Gärtner et al. 2022a,b; Huang et al. 2022; Innmann et al. 2016; Luo et al. 2022; Rempe et al. 2020; Shimada et al. 2021, 2020; Xie et al. 2021; Yuan et al. 2021] or with objects [Dabral et al. 2021]. Some works reconstruct 3D poses from egocentric views [Luo et al. 2021] or IMUs [Yi et al. 2022]. Hu *et al.* [2022] reconstruct hand-object interactions from an RGB-D camera sequence modelling the physics-based contact

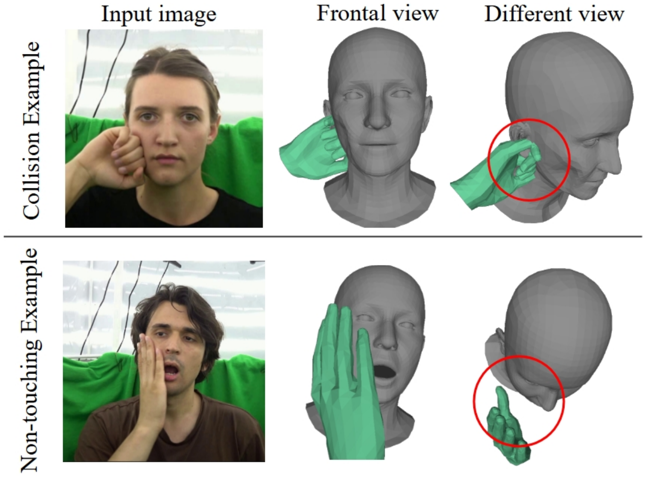

Fig. 3. Example artefacts caused by the depth inaccuracies after solving a naïve single RGB based fitting optimisation, *i.e.* Eqs. (5) and (8) without $\mathcal{L}_{\text{touch}}$, $\mathcal{L}_{\text{col.}}$ and $\mathcal{L}_{\text{depth}}$. The first row shows the physically implausible collisions between the hand and face. The second row displays the "non-touching" artefacts in which no hand-face interactions are discernible in the reconstruction, despite the presence of such interactions in the image input. The locations of the artefacts are indicated by the red circles.

status. While the existing approaches primarily focus on modelling the interactions with static floor planes or rigid objects, our method uniquely addresses non-rigid deformations arising from interactions between hands and face. This capability is made possible thanks to our networks trained on our novel dataset, which incorporates 3D deformations generated using a maker-less multiview motion capture system combined with position based dynamics (PBD) [Müller et al. 2007] – a widely adopted deformable object simulation algorithm employed in modern physics engines.

3 METHOD

Our goal is to reconstruct hands interacting with a face in 3D, including non-rigid face deformations caused by the interaction, from a single monocular RGB video. Figure 2 provides an overview of the proposed framework. Our deformation and contact estimation network *DefConNet*, trained on our new dataset (Sec. 4), estimates face surface deformations and contact labels on both face and hand surfaces from an image sequence; the contact labels are crucial to achieve plausible and realistic interactions in 3D (Sec. 3.2). The estimated deformations, contacts and 2D keypoints are subsequently sent to the global fitting optimisation stage (Sec. 3.3), where we also utilise the *interaction prior* obtained from a conditional variational autoencoder [Sohn et al. 2015] conditioned on the 2D key points for the improved interactions between the hands and face. After this stage, we obtain the final 3D reconstruction of the face and hands in the form of parametric hand and head models with applied deformations. We next explain the notations and assumptions used in this work (Sec. 3.1), followed by the details of our *Decaf* approach.

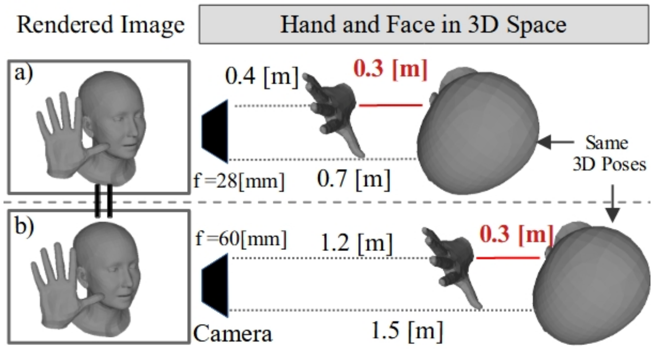

Fig. 4. Schematic visualisation of depth ambiguity in a monocular setup. f denotes the focal length of the camera. **a) and b):** Given the same 3D poses of face and hand of the same scale in the 3D space, different combinations of depths and focal lengths can result in indistinguishable images after the 2D projection in a monocular setting. This effect, known as depth ambiguity, poses a challenge for methods attempting to estimate the depth values of the hand and face in the camera frame from monocular 2D inputs (*e.g.,* RGB images or 2D keypoints). However, the relative location of the hand w.r.t. the head is invariant to the positions of the face and hand in 3D space (*e.g.,* 0.3 [m] above). Based on this idea, our DePriNet learns the depth prior in the **canonical face frame.**

3.1 Modelling and Preliminaries

Our *Decaf* accepts as input a sequence $\mathbf{I} = \{\mathbf{I}_t\} = \{\mathbf{I}_1, ..., \mathbf{I}_T\}$ of $T = 5$ successive RGB frames from a static camera with known intrinsic camera parameters. We resize $\mathbf{I}_t$ to 224×224 pixels after cropping the detected bounding box around the subject's face and hands in each frame. To represent the 3D face, we employ a gender-neutral version of FLAME parametric model $\mathcal{F}$ [Li et al. 2017]. We utilise its identity parameters $\boldsymbol{\beta}_{\text{f}} \in \mathbb{R}^{100}$, jaw pose $\boldsymbol{\theta}_{\text{f}} \in \mathbb{R}^3$ and expression parameters $\boldsymbol{\Psi} \in \mathbb{R}^{50}$ combined with the global translation $\boldsymbol{\tau}_{\text{f}} \in \mathbb{R}^3$ and rotation $\mathbf{r}_{\text{f}} \in \mathbb{R}^3$ that can be formulated as a differentiable function $\mathcal{F}(\boldsymbol{\tau}_{\text{f}}, \mathbf{r}_{\text{f}}, \boldsymbol{\beta}_{\text{f}}, \boldsymbol{\theta}_{\text{f}}, \boldsymbol{\Psi})$. Model $\mathcal{F}$ returns 3D head vertices $\mathbf{V}_{\text{f}} \in \mathbb{R}^{M \times 3}$ (M = 5023) from which we obtain the 3D face landmarks $\mathbf{J}_{\text{f}} \in \mathbb{R}^{K_{\text{f}} \times 3}$ (K_{f} = 68). To represent 3D hands, we employ the gender neutral version of the statistical MANO parametric hand model [Romero et al. 2017] that defines the hand mesh as a function $\mathcal{M}(\boldsymbol{\tau}_{\text{h}}, \mathbf{r}_{\text{h}}, \boldsymbol{\theta}_{\text{h}}, \boldsymbol{\beta}_{\text{h}})$ of global translation $\boldsymbol{\tau}_{\text{h}} \in \mathbb{R}^3$ and global root orientation $\mathbf{r}_{\text{h}} \in \mathbb{R}^3$, pose parameters $\boldsymbol{\theta}_{\text{h}} \in \mathbb{R}^{45}$ and hand identity parameters $\boldsymbol{\beta}_{\text{h}} \in \mathbb{R}^{10}$. This function $\mathcal{M}$ returns hand 3D mesh vertices $\mathbf{V}_{\text{h}} \in \mathbb{R}^{N \times 3}$ (N = 778) from which 3D hand joint positions $\mathbf{J}_{\text{h}} \in \mathbb{R}^{K_{\text{h}} \times 3}$ (K_{h} = 21) are obtained. We assume that the face identity and hand shape parameters are known. In the following, $\boldsymbol{\Phi}_{\text{f}} = (\boldsymbol{\tau}_{\text{f}}, \mathbf{r}_{\text{f}}, \boldsymbol{\beta}_{\text{f}}, \boldsymbol{\theta}_{\text{f}}, \boldsymbol{\Psi})$ and $\boldsymbol{\Phi}_{\text{h}} = (\boldsymbol{\tau}_{\text{h}}, \mathbf{r}_{\text{h}}, \boldsymbol{\beta}_{\text{h}}, \boldsymbol{\theta}_{\text{h}})$ denote the kinematic states of the face and hand in a 3D space.

3.2 Interaction Estimation

We introduce a learning-based approach that estimates plausible interactions in a scene, *i.e.,* the vertex-wise face deformations and contacts on the face and hand surfaces given only single-view RGB images. The approach is trained on our new dataset (Sec. 4).

Our neural network accepts as input an image sequence $\mathbf{I}$ and outputs the deformation on the head model as per-vertex displacements in a camera frame $\mathbf{p} \in \mathbb{R}^{M \times 3}$, contact labels on the face $\mathbf{c}_{\mathrm{f}} \in \{0, 1\}^M$ and the hand $\mathbf{c}_{\mathrm{h}} \in \{0, 1\}^N$. The contact labels are binary signals *i.e.* 1 for contact, 0 otherwise. The network is trained to estimate the contact probability using the binary cross entropy (BCE):

$$\mathcal{L}_{\mathrm{labels}} = \mathrm{BCE}(\mathbf{c}_{\mathrm{f}}, \hat{\mathbf{c}}_{\mathrm{f}}) + \mathrm{BCE}(\mathbf{c}_{\mathrm{h}}, \hat{\mathbf{c}}_{\mathrm{h}}), \tag{1}$$

where $\hat{\mathbf{c}}_{\mathrm{f}}$ and $\hat{\mathbf{c}}_{\mathrm{h}}$ denote the ground-truth contact labels for the face and hand, respectively. We also train the network to estimate the deformations using the ground-truth annotations $\hat{\mathbf{p}}_m$:

$$\mathcal{L}_{\mathrm{def.}} = \frac{1}{M} \sum_{m=1}^{M} (w_{\mathrm{def}}^m \|\mathbf{p}_m - \hat{\mathbf{p}}_m\|_2^2 + b_{\mathrm{def}}^m \|\mathbf{p}_m\|), \tag{2}$$

where

$$w_{\mathrm{def}}^m = \begin{cases} 0.3, & \text{if } \ \|\hat{\mathbf{p}}_m\| = 0, \\ 1.0, & \text{otherwise,} \end{cases} \qquad b_{\mathrm{def}}^m = \begin{cases} 1, & \text{if } \ \|\mathbf{p}_m\| > \psi, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The first term in Eq. (2) allows the network to learn the 3D deformations in our dataset. The weight w_{def} helps to penalise the network predictions more on deforming vertices. We observe that this weighting strategy improves the network precision as the majority of the face vertices have no deformations. The second loss term in Eq. (2) regularises the unnaturally large deformations on the face surface where b_{def} works as a binary label to penalise only the vertices with deformations greater than $\psi = 0.1$ [m].

3.3 Global Fitting Optimisation

Using the estimated deformations $\mathbf{p}$, contact labels $\mathbf{c}_{\mathrm{f}}$ and $\mathbf{c}_{\mathrm{h}}$ and 2D joint keypoints, we obtain the global positions of the face $\mathbf{\Phi}_{\mathrm{f}}$ and hand $\mathbf{\Phi}_{\mathrm{h}}$ in the 3D scene considering their interactions. In this optimisation step, we also update $\mathbf{p}$ to refine and handle the minor collisions. The objective follows:

$$\mathcal{L}_{\mathrm{opt}}(\mathbf{\Phi}_{\mathrm{f}}, \mathbf{\Phi}_{\mathrm{h}}, \mathbf{p}) = \mathcal{L}_{\mathrm{face}} + \mathcal{L}_{\mathrm{hand}}. \tag{4}$$

The fitting loss term of the face model $\mathcal{L}_{\mathrm{face}}$ reads:

$$\mathcal{L}_{\mathrm{face}}(\mathbf{\Phi}_{\mathrm{f}}, \mathbf{p}) = \mathcal{L}_{\mathrm{2D}} + \mathcal{L}_{\mathrm{reg.}}, \tag{5}$$

where $\mathcal{L}_{\mathrm{2D}}$ and $\mathcal{L}_{\mathrm{reg.}}$ are the weights of the 2D reprojection term and regulariser loss term , respectively. Employing the projection function $\Pi(\cdot)$ with the known camera intrinsics, the 2D reprojection loss term is formulated as follows:

$$\mathcal{L}_{\mathrm{2D}} = \frac{1}{M} \sum_{m=1}^{M} w_{\mathrm{conf.}}^m \|\Pi(\mathbf{J}_{\mathrm{f}}^m) - \hat{\mathbf{J}}_{\mathrm{f}}^m\|_2^2, \tag{6}$$

where $\hat{\mathbf{J}}_{\mathrm{f}}^m$ and $w_{\mathrm{conf.}}^m$ are, respectively, the reference 2D face landmarks and the corresponding confidence value obtained by the method of [Bulat and Tzimiropoulos 2017] given the input image. We also minimise the regulariser loss term $\mathcal{L}_{\mathrm{reg.}}$ to introduce the statistical prior for the shape $\boldsymbol{\beta}_{\mathrm{f}}$ and expression $\mathbf{\Psi}$, and temporal smoothness in the motion:

$$\mathcal{L}_{\mathrm{reg.}} = \lambda_{\boldsymbol{\beta}} \|\boldsymbol{\beta}_{\mathrm{f}}\|_2^2 + \lambda_{\mathbf{\Psi}} \|\mathbf{\Psi}\|_2^2 + \lambda_{\dot{\mathbf{V}}} \|\dot{\mathbf{V}}_{\mathrm{f}}\|_2^2 + \lambda_{\ddot{\mathbf{V}}} \|\ddot{\mathbf{V}}_{\mathrm{f}}\|_2^2, \tag{7}$$

where $\dot{\mathbf{V}}_{\mathrm{f}}$ and $\ddot{\mathbf{V}}_{\mathrm{f}}$ denote the velocity and acceleration of the head vertex positions $\mathbf{V}_{\mathrm{f}}$, respectively. $\lambda_{\bullet}$ denotes a weight of the loss term. The objective for the hand fitting $\mathcal{L}_{\mathrm{hand}}$ optimisation includes the 2D reprojection term $\mathcal{L}_{\mathrm{2D}}$, regulariser term $\mathcal{L}_{\mathrm{reg.}}$, collision term $\mathcal{L}_{\mathrm{col.}}$, *touchness* term $\mathcal{L}_{\mathrm{touch}}$ and the depth prior term $\mathcal{L}_{\mathrm{depth}}$:

$$\begin{aligned} \mathcal{L}_{\mathrm{hand}}(\mathbf{\Phi}_{\mathrm{h}}, \mathbf{p}) = \mathcal{L}_{\mathrm{2D}} + \mathcal{L}_{\mathrm{reg.}} + \lambda_{\mathrm{touch}} \mathcal{L}_{\mathrm{touch}} \\ + \lambda_{\mathrm{col.}} \mathcal{L}_{\mathrm{col.}} + \lambda_{\mathrm{depth}} \mathcal{L}_{\mathrm{depth}}, \end{aligned} \tag{8}$$

where $\lambda_{\bullet}$ are the corresponding weights. The terms $\mathcal{L}_{\mathrm{2D}}$ and $\mathcal{L}_{\mathrm{reg.}}$ are the same as in (6)-(7) with the modification that (6) is applied on the hand 3D joints $\mathbf{J}_{\mathrm{h}}$ compared with the reference 2D hand keypoints $\hat{\mathbf{J}}_{\mathrm{h}}$, and (7) on the hand shape $\boldsymbol{\beta}_{\mathrm{h}}$, velocity and acceleration of hand vertices, excluding the expression prior loss $\|\mathbf{\Psi}\|_2^2$.

Due to the inaccuracy of the depth estimation in the monocular setting, simply solving the fitting optimisation w.r.t. the face and hand global positions can cause implausible artefacts, *e.g.* collisions between the face and hand or non-touching artefacts. Figure 3 shows examples of such artefacts, when solving a naïve 2D reprojection based single view fitting optimisation *i.e.* (4) excluding $\mathcal{L}_{\mathrm{touch}}$, $\mathcal{L}_{\mathrm{col.}}$ and $\mathcal{L}_{\mathrm{depth}}$. They immediately give the impression of unnatural hand-face interaction to the viewer. To address the "non-touching" artefacts, we utilise the *touching* loss term $\mathcal{L}_{\mathrm{touch}}$ that penalises the distances between the contact surfaces on the face and hands inspired by [Shimada et al. 2022]. Specifically, we treat the face and hand vertices with contact probabilities $\mathbf{c}_{\mathrm{f}} > 0.5$ and $\mathbf{c}_{\mathrm{h}} > 0.5$ as effective contacts, respectively. Let $C_{\mathrm{f}} \subset [\![1, n]\!]$ and $C_{\mathrm{h}} \subset [\![1, m]\!]$ be the index subsets of the face and hand vertices with the effective contacts. Using a Chamfer loss, $\mathcal{L}_{\mathrm{touch}}$ is formulated as follows:

$$\mathcal{L}_{\mathrm{touch}} = \frac{1}{|C_{\mathrm{f}}|} \sum_{i \in C_{\mathrm{f}}} \min_{j \in C_{\mathrm{h}}} \left\|\mathbf{V}_{\mathrm{f}}^i - \mathbf{V}_{\mathrm{h}}^j\right\|_2^2 + \frac{1}{|C_{\mathrm{h}}|} \sum_{j \in C_{\mathrm{h}}} \min_{i \in C_{\mathrm{f}}} \left\|\mathbf{V}_{\mathrm{f}}^i - \mathbf{V}_{\mathrm{h}}^j\right\|_2^2. \tag{9}$$

To avoid collisions between hands and a head, we also introduce the collision loss term $\mathcal{L}_{\mathrm{col.}}$ for minimising the penetration distance of the hand vertices. Specifically, we first detect the hand vertices colliding with the face mesh based on an SDF criterion [Yu 2023]. Then, we minimise the distance between colliding hand vertices and their nearest vertices on the head mesh. Let $\mathcal{P} \subset [\![1, W]\!]$ be the subset of indices of hand vertices $\mathbf{V}_{\mathrm{h}}$ colliding with the face mesh. The collision loss is formulated as:

$$\mathcal{L}_{\mathrm{col.}} = \sum_{i \in \mathcal{P}} \min_{j \in \mathcal{V}_{\mathrm{f}}} \left\|\mathbf{V}_{\mathrm{h}}^i - \mathbf{V}_{\mathrm{f}}^j\right\|_2^2 + \mathcal{L}_{\mathrm{regDef}}, \tag{10}$$

where $\mathcal{V}_{\mathrm{f}} \subset [\![1, M]\!]$ is the set of all the indices of the face vertices $\mathbf{V}_{\mathrm{f}}$. The term $\mathcal{L}_{\mathrm{regDef}}$ regularises the update of the deformation $\mathbf{p}$ from the perspective of edge lengths, neighbouring face angles and original deformation estimated by DefConNets. Let $l = \{l_1, ..., l_x\}$ and $\varphi = \{\varphi_1, ..., \varphi_y\}$ be vectors that consist of the edge lengths and the angles between the neighbouring faces of the face mesh, respectively. The formulation of $\mathcal{L}_{\mathrm{regDef}}$ reads:

$$\mathcal{L}_{\mathrm{regDef}} = \sum_{i=1}^{x} s_{\mathrm{edge}}^i \|l_i - l_0\|_2^2 + \sum_{i=1}^{y} s_{\mathrm{bend}}^i \|\varphi_i - \varphi_0\|_2^2 + \|\mathbf{p} - \mathbf{p}_0\|_2^2, \tag{11}$$

where l_0 and φ_0 denote the edge lengths and dihedral angles at rest and $\mathbf{p}_0$ is the displacements estimated by DefConNets in the previous step; s_{edge} and s_{bend} are, respectively, the edge and bending stiffness values that consider the underlying skull structure of a human head. The details of the stiffness computations are elaborated in Sec. 4.2.

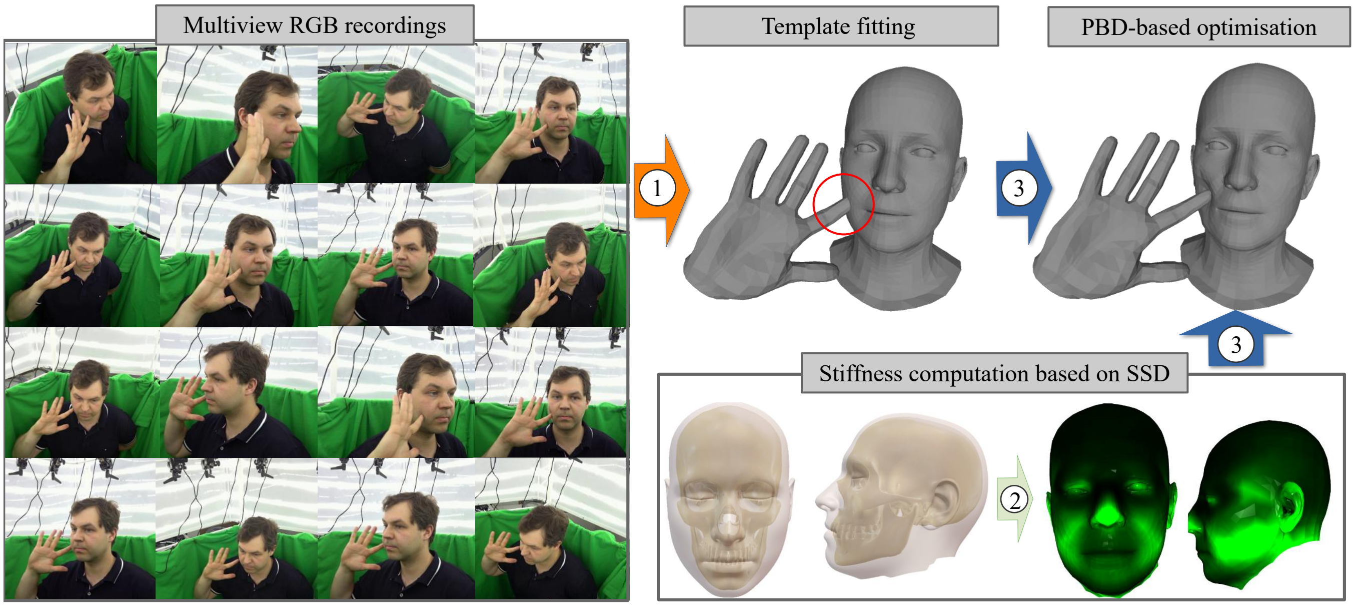

Fig. 5. **Overview of the dataset generation pipeline.** We first capture the hand and face interactions using a markerless multi-view setup. **(1)** Subsequently, the obtained RGB image sequences are used to solve template-based fitting optimisation using MANO [Romero et al. 2017] and FLAME [Li et al. 2017] models. At this stage, due to the unawareness of the hand-face interactions, the collisions are observable, as indicated by the red circle. **(2)** To provide the plausible stiffness values on the head mesh for the later position-based dynamics (PBD) optimisation stage, we compute skull-skin distances (SSD) and obtain vertex-wise stiffness values. (Left-hand side): A visualisation of mean skull and skin surface of a statistic model [Achenbach et al. 2018] from the side and frontal views. (Right-hand side): Transferred stiffness value to FLAME head model [Li et al. 2017] based on the SSD calculation, see Sec. 4.2 for the details. **(3)** Using the fitted templates from (1) and the stiffness values from (2), we solve the PBD-based tracking optimisation. This stage handles the physically implausible collisions and provides plausible surface deformations on the head mesh surface (Sec. 4.3).

To further introduce the learned prior for the depth position of the hand, we train a conditional variational autoencoder (CVAE) [Sohn et al. 2015] -based depth prior network *DePriNet* that is conditioned on the 2D key points. DePriNet is trained to reconstruct the 3D hand key points in a **canonical face frame**, as estimating the depth of hand and face in the camera frame only from monocular 2D input is challenging due to the depth ambiguity (*e.g.* 3D hand and face with different combinations of focal lengths and depths can be projected onto the same position in the 2D image). However, the hand positions relative to the face in the 3D space are invariant to the depth in the camera frame; see Fig. 4 for a schematic visualisation. We train DePriNet with the standard losses:

$$\mathcal{L}_{\text{vae}} = \left\|\mathbf{J}^*_{\text{h}} - \hat{\mathbf{J}}^*_{\text{h}}\right\|_2^2 + \text{KL}\left(q(\mathbf{Z} \mid \hat{\mathbf{J}}^*_{\text{h}}, \mathbf{\Theta}) \| \mathcal{N}(\mathbf{0},\mathbf{I})\right). \tag{12}$$

The first term is a reconstruction loss to reproduce the ground-truth input hand joints in a canonical face frame $\hat{\mathbf{J}}^*_{\text{h}} \in \mathbb{R}^{K_{\text{h}}\times 3}$ and $\mathbf{J}^*_{\text{h}} \in \mathbb{R}^{K_{\text{h}}\times 3}$ denotes the output from the decoder network of DePriNet. The second loss term penalises the deviation of the latent vector $\mathbf{Z} \in \mathbb{R}^{50}$ distribution from a standard normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$ using the Kullback-Leibler divergence loss $\text{KL}(\cdot\|\cdot)$. Latent $\mathbf{Z}$ is sampled from a Gaussian distribution whose mean and variance are estimated from the encoder network $q(\cdot)$ of DePriNet. At test time, we use the decoder network $p(\cdot)$ of DePriNet to output depth candidates of the hand positions that are integrated into the depth prior loss $\mathcal{L}_{\text{depth}}$ in the global fitting optimisation:

$$\mathcal{L}_{\text{depth}} = \sum_{i=1}^{u} w_i \left\|\mathbf{J}^z_{\text{h}} - \mathbf{T}(\mathbf{J}^*_{\text{h},i})\right\|_2^2, \tag{13}$$

$$\text{where } w_i = 1 - \frac{\eta_i - \min(\eta)}{\max(\eta) - \min(\eta)}, \quad \eta_i = |\mathbf{Z}^i|_1, \tag{14}$$

$\mathbf{J}^z_{\text{h}}$ denotes the z-value of the hand 3D keypoints $\mathbf{J}_{\text{h}}$ that corresponds to the depth axis in the camera frame, and $\mathbf{T}(\cdot)$ is a transformation from the canonical face space to the camera frame that consists of the rotation and translation of the face model (that are also simultaneously obtained in this global fitting optimisation); $\mathbf{J}^*_{\text{h},i}$ is the i-th sample obtained from the decoder $p(\cdot)$ given $u = 100$ latent vectors $\sim\mathcal{N}(\mathbf{0},\mathbf{I})$ and the conditioning vector $\mathbf{\Theta}$ that consists of face and hand 2D keypoints with corresponding confidence values as well as the face 3D rotation in the camera frame in 6D representation [Zhou et al. 2019]. Note that 2D key points of the face and hands are translated to be a face-root relative representation for the conditioning. The conditioning 3D head rotation is obtained during the optimisation (4). Each generated sample is weighted by the scalar w that has the higher value the closer the corresponding latent vector $\mathbf{Z}$ is to zero (*i.e.* a statistically more likely sample). We utilise the two independent DePriNets of the same architecture for the left and right hands. After minimising the objective that combines all these loss terms, we obtain the final 3D head and hand reconstructions

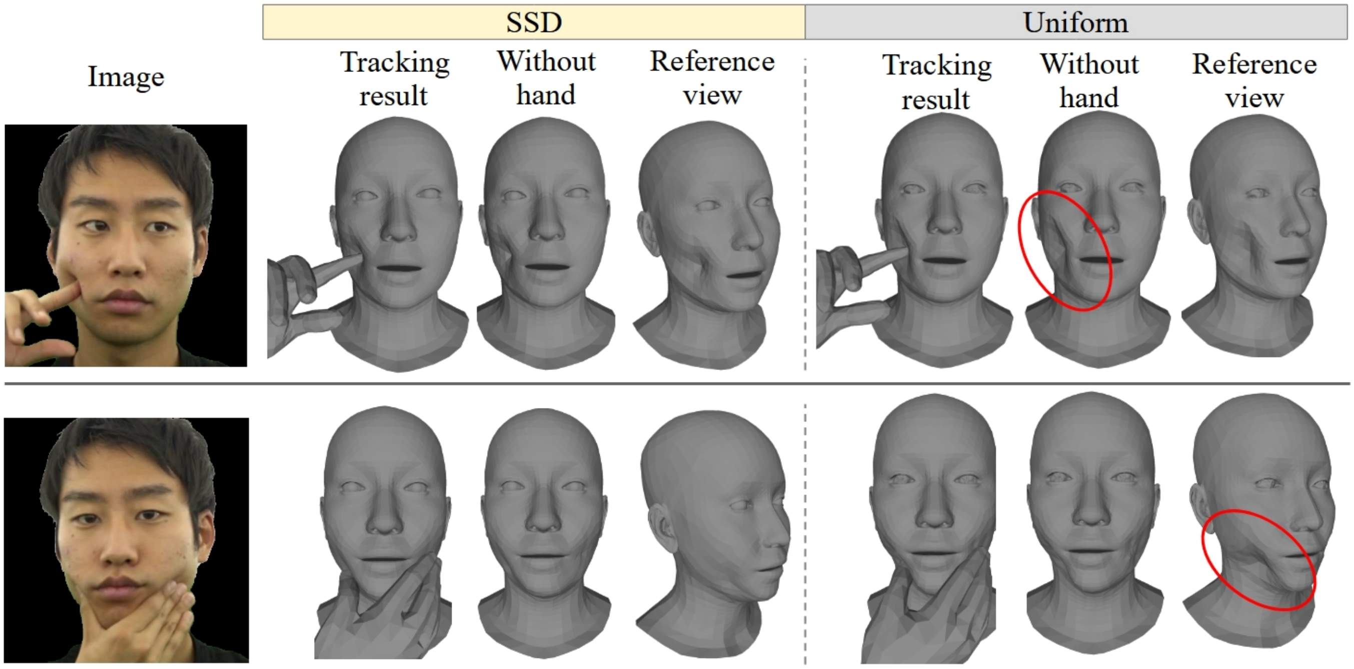

Fig. 6. Example visualisations of the reconstructed 3D head and hand interactions with the stiffness values computed using the skull-skin distance (SSD) (second to fourth columns) and the uniform stiffness value (fifth to seventh columns). With SSD, the obtained surface deformations are much more plausible compared to naïevely assigning the uniform stiffness value to all the head vertices. The red circles highlight the overly deformed surfaces (left) and inaccurate deformations that ignore the underlying jaw in the human head (right).

with plausible deformations and interactions. The significance of each loss term is evaluated in Sec. 5. The final deformed face vertices $\mathbf{V}_\mathrm{f}^*$ are obtained by simply adding the updated deformations $\mathbf{p}$ to the face model parameterised by $\mathbf{\Phi}_\mathrm{f}$, *i.e.* $\mathbf{V}_\mathrm{f}^* = \mathcal{F}(\mathbf{\Phi}_\mathrm{f}) + \mathbf{p}$.

3.4 Architectures of Our Networks

Our *Decaf* comprises several components (Fig. 2). We employ [Bulat and Tzimiropoulos 2017] and [Lugaresi et al. 2019] for 2D keypoint and bounding box estimation of the face and hand, respectively. The *DefConNet* is composed of two encoders and three decoders. The encoders for the cropped face and hand images follow the ResNet-18 architecture [He et al. 2016]. The decoders, sharing the same architecture, estimate per-vertex deformations and contact labels for the face and hand. Each of them includes three fully connected layers with leaky ReLU activation [Maas et al. 2013] and their hidden layer dimensions equal to 1024. We duplicate *DefConNet* for both hands and compute the union of the face deformations and contacts before the final global fitting optimisation. The *DePriNet* is a variational autoencoder [Kingma and Welling 2014], consisting of three linear fully connected layers with batch normalisations, ReLU activations [Agarap 2018], a latent dimension of 50 and hidden size of 128 for both encoders and decoders.

4 DATASET

In this work, we build a new markerless multi-view dataset for 3D hand-face interactions for method training and evaluation. It contains eight subjects—captured with 15 SONY DSC-RX0 cameras at 50 fps (*i.e.,* from 15 different viewpoints)—along with the corresponding reference 3D geometries of a right hand and head, including surface deformations of the head represented as per-vertex displacements. In total, the dataset contains 100K frames, see Table 1 for the details. Each actor performs seven different actions with three different facial expressions. For each captured view, the background masks are obtained using [Sengupta et al. 2020]. The bounding boxes (for the hands and the faces) and 2D key points (for the faces), are obtained using [Lugaresi et al. 2019] and [Bulat and Tzimiropoulos 2017], respectively.

In the remainder of this section, we elaborate on our dataset generation pipeline; see Fig. 5 for the overview. The first step of the pipeline, *i.e.,* multiview template fitting, is explained In Sec. 4.1. Next, to obtain a reasonable stiffness value that considers the underlying skull structure of a human face, we introduce a simple but effective skull-skin distance (SSD) approach in Sec. 4.2. The computed stiffness values are further utilised in the deformable object simulation relying on *position based dynamics (PBD)*, and we obtain the final 3D geometry with plausible interactions arising from hand-face interactions (Sec.4.3).

4.1 Multiview Template Fitting

We first solve the 2D keypoint reprojection-based fitting optimisation to obtain the MANO [Romero et al. 2017] and FLAME model [Li et al. 2017] parameters, so that the hand and face shapes match the multiview 2D keypoints with known intrinsic and extrinsic calibrations. The objective for the face fitting encompasses (6) and (7). For

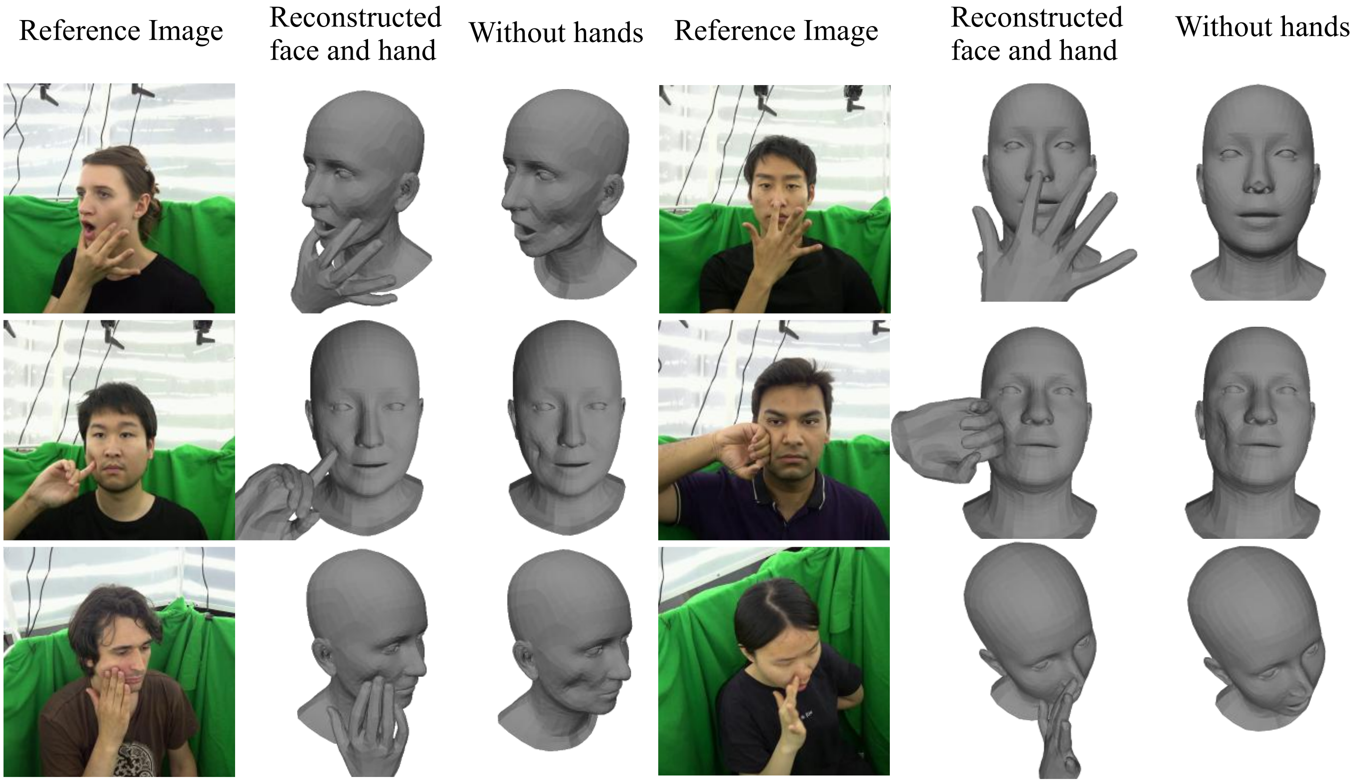

Fig. 7. Example visualisations from our new hands+face 3D motion capture dataset with hand shape articulations non-rigid face deformation. The reconstructed 3D geometry shows plausible surface deformations thanks to the fitting optimisation combined with PBD.

the hand, we also minimise (6) and (7) with the modification that (6) is applied on the hand 3D joints $\mathbf{J}_\mathrm{h}$, and (7) is applied on the hand shape $\boldsymbol{\beta}_\mathrm{h}$, velocity and acceleration of hand vertices, excluding the expression loss term $\left\|\Psi\right\|_2^2$. However, FLAME does not model the surface deformation caused by the interactions, which can result in physically implausible collisions; see the red circle in Fig. 5-(1). We address this limitation by integrating into our tracking pipeline a deformable object simulator relying on position-based dynamics (PBD) [Müller et al. 2007]. Our approach assumes non-homogeneous stiffness values of the human face, and we describe next how we obtain those.

4.2 Stiffness on a Head Mesh

Deformable object simulators require known material stiffness. The stiffness of human face tissues is non-uniform, due to the rich mimic musculature and the skull anatomy. Therefore, assuming uniform stiffness in the whole face and head would result in physically implausible artefacts when running the simulation; see Fig. 6 for the examples. We obtain the non-uniform stiffness values based on a simple but effective *skin-skull distance (SSD)* assumption. It is based on the assumption that our face and head region tend to have higher stiffness when the distance between the skin and skull surface is smaller (*e.g.* forehead), and vice versa (*e.g.* cheek). To compute SSD, we employ the mean skull and skin surface of a statistic model from [Achenbach et al. 2018]. The obtained tissue stiffness map is is upon our expectation and the corresponding pseudo-ground-truth deformations are used in quantitative experiments in Sec. 5.

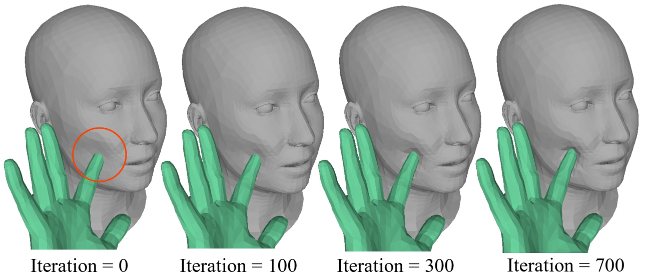

Fig. 8. Visualisation of the effect of $\mathcal{L}_{\mathrm{col.}}$ (10). Starting from the colliding hand and face poses (left-most visualisation), our non-rigid collision loss term effectively resolves the physically implausible inter-penetrations in the course of the optimisation.

Let $\mathbf{D} = [d_1, ..., d_h] \in \mathbb{R}^h$ be a set of nearest distances between the skin and skull surfaces computed for all the h skin vertices of [Achenbach et al. 2018]. The stiffness s of the i-th skin vertex is calculated as follows:

$$\mathbf{s}_i = (1 - \hat{d}_i)^b, \tag{15}$$

Table 1. Details of our new dataset. This dataset contains several types of data including pseudo ground truth of 3D surface deformations represented as 3D displacement vectors for seven different actions with three different facial expressions performed by eight subjects. The "Age" signifies the age range, whereas the number in the brackets means the corresponding number of subjects.

Characteristic	Value/Description
Number of subjects	8
Number of views	16
Total Number of Frames	100 K
Ethnicity	5 Asian, 3 Caucasian
Gender	6 male, 2 female
Age	20 - 29 (5), 30 - 39 (3)
Facial expressions	neutral, open mouth, smiling
Action types	poking a cheek (open hand) poking a cheek (pointing hand) punching a cheek pushing a cheek with a palm rubbing a cheek pinching a chin touching nose front touching nose from side
Data types	2D hand keypoints 2D face landmarks RGB videos foreground segmentation masks hand-face bounding box 3D mesh for hand and face 3D surface deformations

where $\hat{d}$ is the normalised distance:

$$\hat{d}_i = \frac{d_i - \min(\mathbf{D})}{\max(\mathbf{D}) - \min(\mathbf{D})}, \tag{16}$$

with the operators $\min(\cdot)$ and $\max(\cdot)$ to compute the minimum and maximum values of the input vector; b is empirically set to 4. After computing the per-point stiffness $\mathbf{s}_i$, we transfer it to the FLAME head model by finding the corresponding vertices based on the nearest neighbour search after fitting the FLAME head model onto the skin surface model of [Achenbach et al. 2018]. In Fig. 5-(2), we show the visualisation of the assigned stiffness values (more saturated green encodes lower stiffness). The assigned values are expected from the anatomical viewpoint (*e.g.* high stiffness around the head region and low stiffness near the tip of the nose and cheeks). The edge and bending stiffness values in (11) are obtained by simply computing the average over the s of vertices that form the edges and triangles.

4.3 PBD-based Optimisation

Position based dynamics (PBD) [Müller et al. 2007] is a technique for simulating deformable objects, which gained popularity for its robustness and simplicity; it is widely used in game and physics engines. We utilise PBD to resolve implausible head-hand collisions which are challenging to address in a markerless motion capture setup due to constant occlusions at the interaction regions. We utilise stretch constraint C_{stretch}, bending constraint C_{bend} and collision constraint $C_{\text{collision}}$ in the PBD simulator. For each pair of connected vertices $\mathbf{p}_1$ and $\mathbf{p}_2$ in the mesh, C_{stretch} is defined as follows:

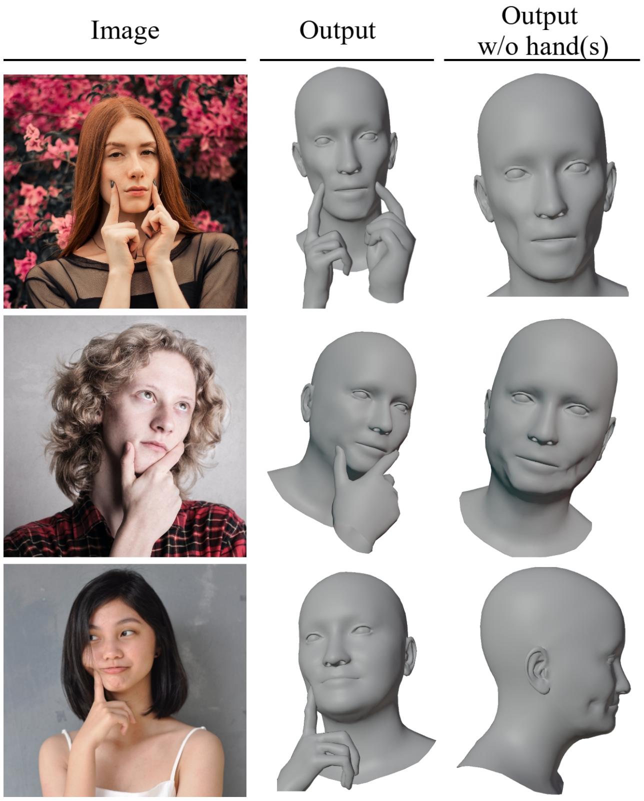

Fig. 9. 3D reconstructions on unseen identities in the wild. Our *Decaf* reasonably generalises across different identities and illuminations unseen during the training. The images are taken from [Pexels 2023].

$$C_{\text{stretch}}(\mathbf{p}_1, \mathbf{p}_2) = |\mathbf{p}_1 - \mathbf{p}_2| - l_0, \tag{17}$$

where l_0 denotes the rest length of the edge between $\mathbf{p}_1$ and $\mathbf{p}_2$. For each pair of adjacent triangles $(\mathbf{p}_1, \mathbf{p}_3, \mathbf{p}_2)$ and $(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_4)$, the definition of bending constraint C_{bend} reads:

$$C_{\text{bend}}(\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \mathbf{p}_4) = \\ \operatorname{acos}\left(\frac{(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_3 - \mathbf{p}_1)}{|(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_3 - \mathbf{p}_1)|} \cdot \frac{(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_4 - \mathbf{p}_1)}{|(\mathbf{p}_2 - \mathbf{p}_1) \times (\mathbf{p}_4 - \mathbf{p}_1)|} \right) - \varphi_0, \tag{18}$$

where φ_0 is the rest angle between the two triangles. Collision constraint $C_{\text{collision}}$ can be integrated for each vertex $\mathbf{p}$:

$$C_{\text{collision}}(\mathbf{p}) = \mathbf{n}^T \mathbf{p} - h = 0, \tag{19}$$

where $\mathbf{n}$ and h are the normal of the colliding plane and the distance from the plane that $\mathbf{p}$ should maintain. After resolving collisions, we introduce friction as formulated in [Müller et al. 2007] with 0.5 for both kinetic and static friction coefficients.

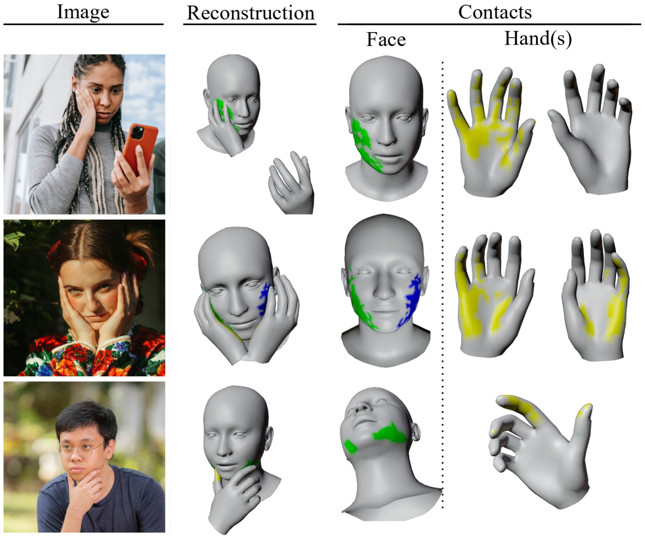

Fig. 10. Visualisations of the estimated contacts on in-the-wild images. The green and blue colours represent the face contacts regressed by the right- and left-hand DefConNet, respectively (see Fig. 2). The yellow colour represents the contact regions on the hand(s). All estimations are reasonable. The images are taken from [Pexels 2023].

We also additionally introduce constraint C_{track} for tracking the reference 3D motions obtained in Sec. 4.1. More specifically, this tracking constraint minimises the Euclidean distance between the vertex of the template mesh $\mathbf{p}$ and its corresponding vertex $\mathbf{p}_{\text{ref}}$ in the reference mesh from the previous multi-view fitting stage:

$$C_{\text{stretch}}\ (\mathbf{p}, \mathbf{p}_{\text{ref}}) = |\mathbf{p} - \mathbf{p}_{\text{ref}}|\,. \tag{20}$$

For the simulation, we use the stiffness values obtained in Sec.4.2, and finally obtain the 3D geometry of the interacting hand and face with the surface deformations (also see Fig. 5-(3) for the example reconstruction).

5 EVALUATIONS

We next evaluate our *Decaf* on our new dataset. As there are no existing methods that address the same problem we tackle, we compare our method to a most closely related approach, *i.e.,* a monocular full-body capture PIXIE [Feng et al. 2021a] and its variants that reconstruct only hands and face independently, denoted as PIXIE (hand+face). We also compare to our benchmark method that includes hand-only [Lugaresi et al. 2019] and face-only [Li et al. 2017] trackers.

Note that in this method variant, DefConNet and non-rigid collision handling (10) are deactivated. Our dataset contains separate training and testing sequences containing the same kinds of actions. We train our networks on the training sequences of 5 different subjects and conduct the quantitative evaluations on 3 different subjects unseen during the training. For the qualitative comparisons, we show the results of our data recording green studio and indoor sequences captured using a SONY DSC-RX0 camera.

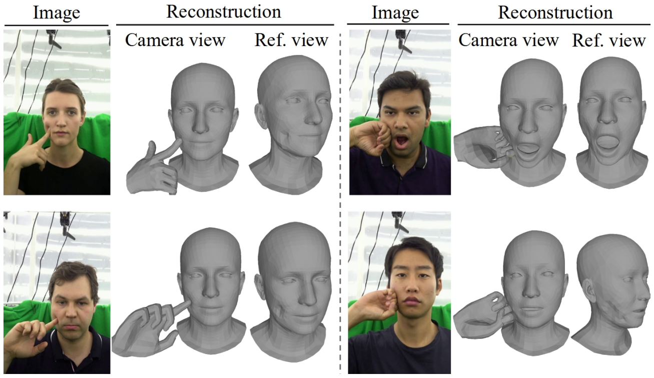

Fig. 11. 3D reconstructions on actions unseen during the training, *i.e.* (left:) poking a cheek (pointing hand) and (right:) punching a cheek.

5.1 Implementation and Training Details

The neural networks were implemented in PyTorch [Paszke et al. 2019]. The evaluations and network training were conducted on a computer with an NVIDIA QUADRO RTX 8000 graphics card and AMD EPYC 7502P 32 Core Processor. The training was continued until convergence using Adam optimiser [Kingma and Ba 2014] with a learning rate $3 \cdot 10^{-4}$. *DefConNet* models are trained until convergence which takes ≈12 hours. Since our dataset was captured with right-hand and face interactions, we flip the image and the corresponding 3D ground-truth annotations and contact labels horizontally to obtain the input and ground truth for the left hand. For the global fitting optimisation, we set the loss term weights of (5), $\lambda_{\boldsymbol{\beta}} = 1 \cdot 10^{-5}$, $\lambda_{\boldsymbol{\Psi}} = 1 \cdot 10^{-3}$, $\lambda_{\dot{\mathbf{V}}} = 3 \cdot 10^{-4}$, $\lambda_{\ddot{\mathbf{V}}} = 3 \cdot 10^{-4}$. For (8), we employed the following weights: $\lambda_{\text{touch}} = 0.1, \lambda_{\text{col.}} = 1.0$, $\lambda_{\text{depth}} = 3 \cdot 10^{-3}$, $\lambda_{\boldsymbol{\beta}} = 1 \cdot 10^{-5}$, $\lambda_{\dot{\mathbf{V}}} = 3 \cdot 10^{-4}$, $\lambda_{\ddot{\mathbf{V}}} = 3 \cdot 10^{-4}$. As the 2D hand keypoint estimator [Lugaresi et al. 2019] in our method estimates 3D hand key points as well, we utilise them to initialise our hand pose by simply fitting the MANO hand model onto the 3D keypoints using inverse kinematics (Note that this step is optional.).

5.2 Qualitative Evaluations

Our supplementary video shows comparisons of our results with results of PIXIE (hand+face) [Feng et al. 2021a] as well as the benchmark methods in a studio and an indoor scene, *i.e.* monocular hand [Lugaresi et al. 2019] and face [Li et al. 2017] trackers operating independently. Only our method reconstructs face deformations caused by the interactions while showing much more accurate 3D localisations of the hands and face compared to other approaches; see Fig. 13 and Fig. 14 for the visualisations. In Fig. 8, we also show an example visualisation of the non-rigid collision loss (10) starting from colliding hand and face positions. While the optimisation progresses, the physically implausible collisions are resolved by plausibly deforming the face surface. Our qualitative results confirm that *Decaf* produces significantly more plausible hand-face interactions and natural face deformations from a single RGB video compared with others.

To assess the generalisability of our *Decaf* across diverse identities and lighting conditions, we evaluate it on in-the-wild images; see

Table 2. Comparisons of the 3D reconstruction accuracy and plausibility of interactions. Lower PVE indicates higher 3D reconstruction accuracy. "†" denotes PVE after applying a translation on both the face and hand that translates the centre of the face mesh to the origin. Our *Decaf* shows the lowest error in PVE and DefE metrics. In the plausibility measurements, lower Col. Dist. and higher Non. Col. indicate the lower magnitude of collisions and less frequent collisions, respectively (thus, more plausible interactions). Higher Touchness represents higher plausibility of the interaction that corresponds to the image input.

	3D Error		Plausibility Measurement			
	PVE [mm]↓	PVE† [mm]↓	Col. Dist. [mm]↓	Non. Col. [%]↑	Touchness [%]↑	F-Score [%]↑
Ours	**11.9**	**9.65**	1.03	83.6	96.6	**89.6**
Ours w/o $\mathcal{L}_{\text{touch}}$	17.4	15.2	6.83	68.7	78.5	73.2
Ours w/o $\mathcal{L}_{\text{col.}}$	15.7	12.9	14.4	59.6	87.7	71.0
Ours w/o $\mathcal{L}_{\text{depth}}$	15.9	13.8	11.0	77.2	85.5	81.1
Benchmark	18.9	17.7	19.3	64.2	73.2	68.4
PIXIE (hand+face)	41.6	26.3	7.04	75.9	75.1	75.5
PIXIE	51.9	39.7	0.11	97.1	51.8	67.6

Table 3. 3D deformation error comparisons. Lower DefE indicates higher 3D accuracy of the deformations. "+" indicates that DefE was computed only on deformations whose ground-truth deformation vector has a norm greater than 5 [mm]. Our full method shows the lowest deformation error. Note that DefE and +DefE for related methods and benchmarks are computed using zero displacements as only our method outputs the per-vertex deformations (denoted with "*") .

	DefE. [mm]↓	+DefE. [mm]↓
Ours	**0.08**	**2.28**
Ours w/o refinement	0.09	2.35
Benchmark	0.13^*	7.28^*
PIXIE (hand+face)	0.13^*	7.28^*
PIXIE	0.13^*	7.28^*

Fig. 9. The reconstructed 3D shapes show plausible interactions with reasonable facial deformations. Furthermore, the estimated contacts showcased in Fig. 10 faithfully mirror the contact regions evident in the input images. As a result, the final reconstructions show plausible hand-to-face interactions guided by the estimated contacts. To further assess the generalisability of our method on unseen actions, we train our networks excluding "poking a cheek (pointing hand)" and "punching a cheek" actions from the training dataset; the results for these actions are illustrated in Fig. 11. Our method produces satisfactory results for "poking a cheek (pointing hand)". On the other hand, the exclusion of "punching a cheek" from the training dataset is a highly challenging scenario as no other actions in the training data contain interactions between the back side of the hand and the face. Given that our approach is neural and learning-based, such a substantial deviation from the training set can lead to inaccurate interactions in the results.

5.3 Quantitative Evaluations

To evaluate our algorithm from various perspectives numerically, we report multiple evaluation metrics. We calculate the 3D per vertex error (PVE) as an indicator of the 3D accuracy as well as the 3D deformation errors for our estimated face deformations. Additionally, we report the metrics of *collision distance*, *non-collision ratio* and *touchness ratio* to quantify the physical plausibility of the reconstructed hands and faces. We also include the *F-Score* to evaluate the overall plausibility of the reconstructions, taking into account both the occurrences of collisions and the correctness of the interactions. The specific details of each metric are elaborated as follows:

- **Per vertex error (PVE)** measures the magnitude of the 3D error by computing the average Euclidean distances between the reconstruction and the ground-truth vertices. We report the errors in the camera frame before and after applying a translation on the hand and face that aligns the centroid of the face with the origin of the coordinate frame, denoted as PVE and PVE†, respectively. Thus, PVE† measures the reconstruction quality focusing on the relative position of the hand w.r.t. the head, which is important when judging the accuracy of the interactions.
- **Deformation error (DefE)** measures the magnitude of the error by computing the average Euclidean distances between the estimated per-vertex 3D deformations and their pseudo ground truth. We also report **+DefE** that computes DefE only for deformations with the corresponding ground-truth deformation vectors of norm greater than 5 [mm], *i.e.* when non-negligible interactions are present. Lower DefE and +DefE indicate higher prediction accuracy of the deformations.
- **Collision distance (Col. Dist.)** measures the collision distances averaged over the number of vertices and frames. A lower collision distance indicates a smaller magnitude of collisions throughout the sequence.
- **Non-collision ratio (Non. Col.)** measures the ratio of the frames with no collisions between the hand and face over all sequence frames. A higher non-collision ratio indicates fewer collisions in the reconstructed sequence.
- **Touchness ratio** measures the ratio of frames over all the frames where contacts between face and hand are present in the prediction when there are face-hand contacts in our ground truth. The hand vertices with the nearest distance from the face surface lower than 5 [mm] are considered in contact. This metric exposes the presence of an artefact, namely the occurrence of face-hand interactions in the input frame while the hand does

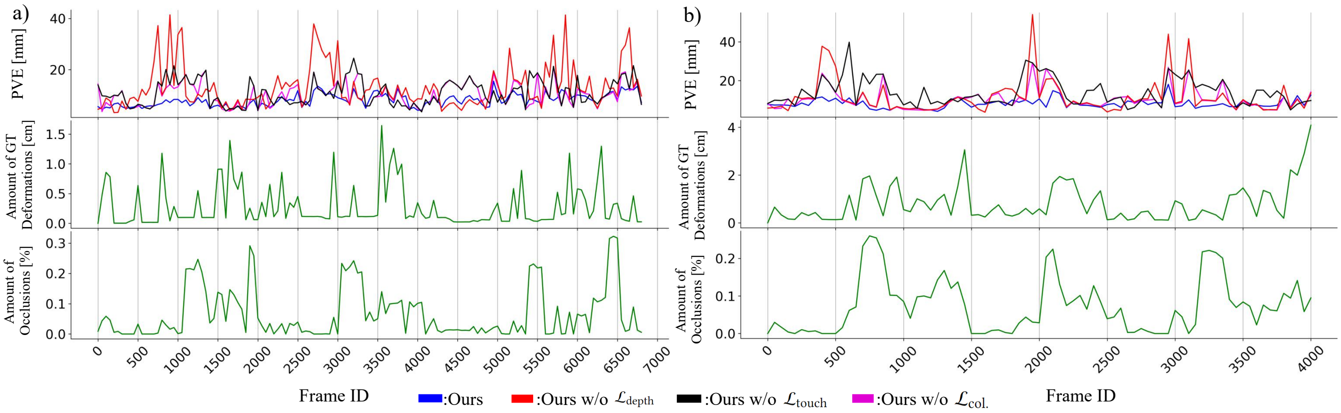

Fig. 12. PVE plots for two exemplary test sequences (left: woman on top-left in Fig. 7; right: man on middle-right in Fig. 7) in relation to the degree of occlusions and deformations in the pseudo ground truth. Our full model is affected by the occlusions (the bottom row) substantially less than its ablated versions.

not make physical contact with the face in the reconstruction. A higher ratio indicates more plausible reconstructions.

- **F-Score** for Non. Col. and touchness ratio are also reported by computing the harmonic mean of the two (as these two metrics are complementary to each other). It is very important to report F-Score, since each of these metrics in isolation is not meaningful (*e.g.* constant presence of hand-face collisions will result in perfect touchness ratio 100%; no presence of interaction throughout the sequence will make the perfect Non. Col. 100%). A higher F-Score indicates a higher plausibility of the interactions in the reconstructions showing fewer occurrences of collisions and incorrect interactions.

3D Error Comparisons. We report PVE in Table 2-(left) to evaluate the 3D accuracy of the reconstructed hand and face. Our *Decaf* shows the best performance scoring around 40% less error compared with the second best method, benchmark ([Lugaresi et al. 2019] + [Li et al. 2017]). We also report the 3D accuracies of the deformations; DefE and +DefE in Table 3. To compute DefE for the related works, we simply provide zero deformations, as those methods do not model per-vertex deformations caused by interactions. For both DefE and +DefE, our method shows the lowest errors, *i.e.* about 60% lower errors for DefE and 40% lower errors compared with others.

Plausibility of Interactions. In Table 2, we report Col. Dist., Non. Col., Touchness and F-Score. It is very important to show F-Score as Non. Col. and Touchness are *complementary to each other*. Ours show low collision distances while showing quite high *Touchness*, which indicates the highly plausible face-hand interactions that correspond to the input images, thus the best performance in F-Score. In contrast, PIXIE shows extremely low collision distances while showing much worse *Touchness* compared with ours. This is because, in most cases, the reconstructed hand and face are wrongly not interacting with each other when they should be interacting; see Fig. 13 for the example reconstructions. The benchmark and PIXIE (hand+face) independently reconstruct the face and hands being agnostic of the interactions of those, therefore they show quite frequent collisions (high Col. Dist. and low Non. Col.) as well as incorrect interactions (Low Touchness), thus lower F-Score than ours. Given these metrics in Table 2 and qualitative results in our video, *Decaf* shows the most plausible interactions in the reconstructed results compared with the related methods.

Table 4. Perfomance measurement of our contact estimation component. Our method estimates reasonable contacts on face-hand surfaces only from RGB input, which are integrated into the final global fitting optimisation. The significance of the contacts is validated in Table 2.

	F-score ↑	Precision ↑	Recall ↑	Accuracy ↑
face	0.57	0.69	0.49	0.99
hand	0.47	0.62	0.39	0.98

Ablation Studies. In Table 2, we show the ablation studies of the reconstructions denoted as “Ours w/o $\mathcal{L}_{\text{touch}}$”, “Ours w/o $\mathcal{L}_{\text{col.}}$ ” and “Ours w/o $\mathcal{L}_{\text{depth}}$ ” to assess the importance of each loss term. For both the 3D accuracy and plausibility measurements, removing one loss term results in a severe performance decrease, which confirms all those loss terms contribute to higher 3D localisations and improvement of interaction plausibilities. Additionally, in Table 3, we also show the DefE and +DefE without updating the deformations in the final global fitting optimisation stage *i.e.* direct output from the DefConNet denoted as “Ours w/o refinement”. Our final global fitting optimisation improves the estimated deformations from DefConNet, reducing the DefE and +DefE by 10% and 3%.

Fig. 12 shows PVE plots for two test sequences from our dataset highlighting the stability of our results. *Amount of occlusion* stands for the per-frame ratio of face pixels occluded by hand pixels from the camera view and *amount of deformations* signifies the per-frame sums of deformations in the pseudo ground truth. We observe that the ablated versions of our method are starkly influenced by occlusions, which can be recognised with the help of peaks occurring at the frames with the (locally) largest deformations as well as the most significant occlusions. In contrast, our full model is affected by the occlusions substantially less and its curve has a smaller standard deviation of PVE, which verifies the importance of each loss term.

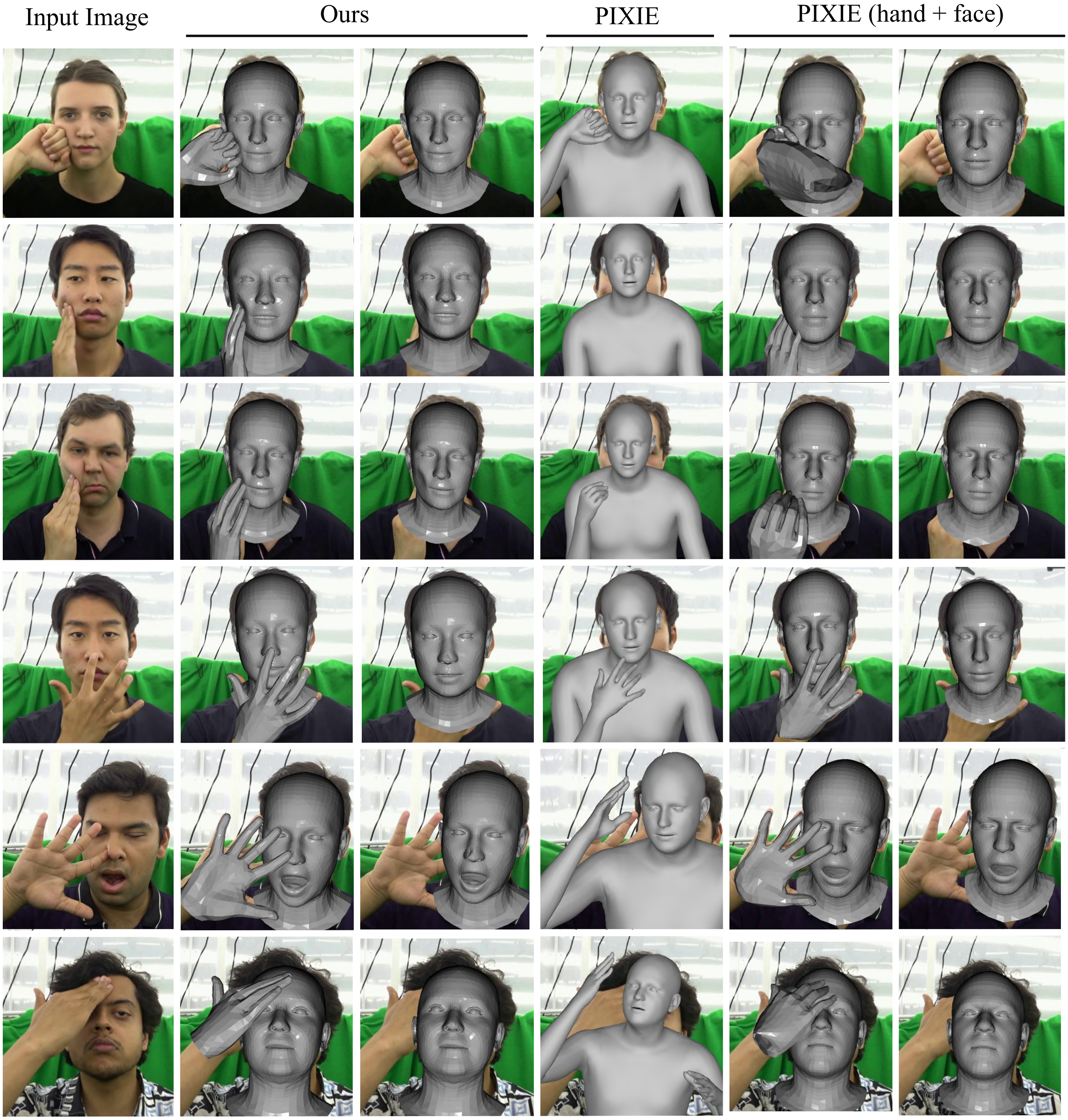

Fig. 13. Visualisations of the experimental results by our method, PIXIE [Feng et al. 2021a] and hand-face only mode of PIXIE. The PIXIE results (fourth column) frequently lack interactions between the hand and face, resulting in a low touchness ratio (Table 2). PIXIE (hand+face) in the fifth column shows collisions and lacks face-hand interactions as the method is agnostic to the latter. Our results (second column) exhibit natural interactions between the hand and face along with plausible face deformations (third column), which are not present in the results of the competing approaches (fourth and sixth columns).

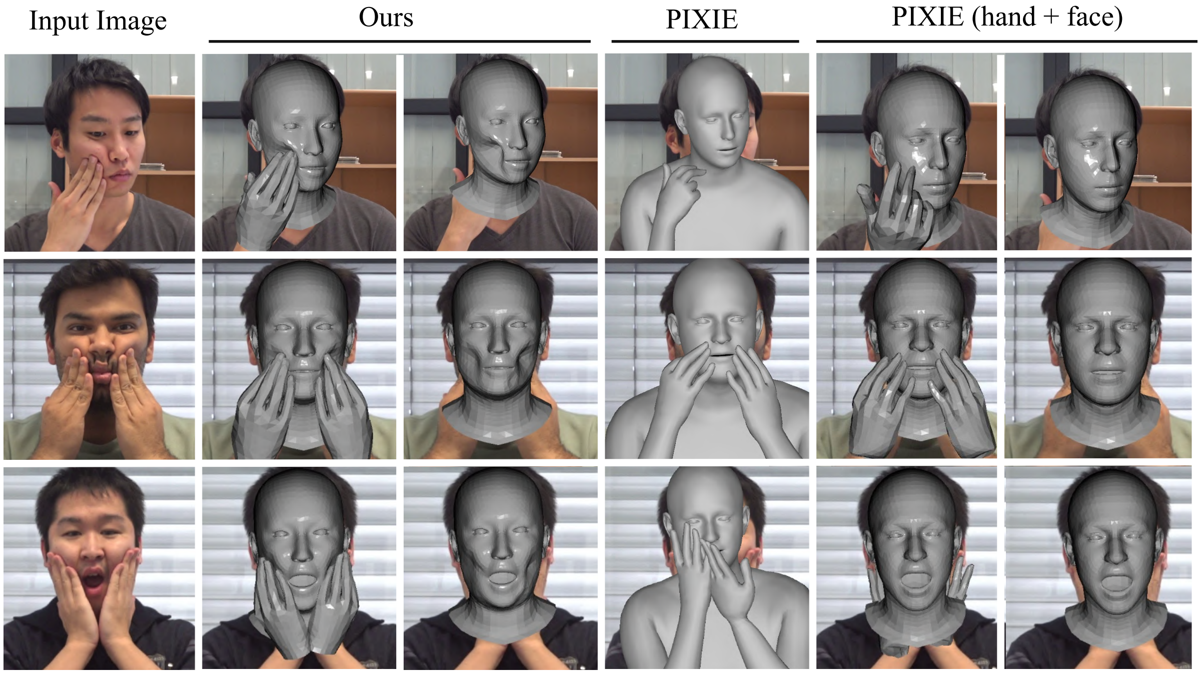

Fig. 14. Visualisations of the experimental results by our method, PIXIE [Feng et al. 2021a] and hand-face-only mode of PIXIE for indoor scenes. Our results are plausible and represent expressive facial deformations, whereas the other works show inaccurate interactions and lack deformations.

Contact Estimations. To our knowledge, there are no existing works that estimate the contacts on hand-face surfaces from RGB inputs. Nonetheless, we report the performance of the contact estimation of our method for comparison on Table 4. Note that although estimating contact vertices only from RGB inputs is a highly challenging problem, our *Decaf* estimates reasonable contacts that significantly improve the 3D localisation as validated in Table 2.

6 DISCUSSIONS AND LIMITATIONS

Our *Decaf* captures plausible 3D deformations along with hand-face interactions solely from a monocular RGB video, effectively reducing unnatural collisions and non-touching artefacts. While our method is the first to address this problem set, it does have certain limitations. Our network learns from a newly created dataset computed using Position-Based Dynamics (PBD) with a skull-skin-distance (SSD) approach combined with the multi-view markerless motion capture setup. PBD is widely utilised in modern physics engines, ensuring that our pseudo-ground truth deformations are plausible. However, it may introduce some discrepancies between the actual deformations and calculated deformations as this PBD-based approach does not integrate visual information such as photometric loss. Nevertheless, we believe this approach to be satisfactorily accurate to obtain plausible deformations although the visual information is not reliable at the interaction regions due to the constant occlusions, which is verified in our qualitative experiments. Our method employs PCA-based parametric face and hand models. Consequently, the 3D reconstructions of both body parts maintain consistent topology though, as a downside, miss high-frequency details such as wrinkles or blood vessels. Lastly, our method primarily focuses on handling pushing actions (*e.g.* pushing or poking cheeks). Furthermore, it is important to note that object-hand-face interactions, which fall outside the scope of our research, can be addressed in future studies.

7 CONCLUSIONS

Decaf is the first monocular RGB-based approach for deformation-aware 3D hand-face motion capture. Our method captures non-rigid face surface deformations arising from various hand-head interactions. It regards the human head anatomy (*i.e.,* skull-skin distance used to calculate non-uniform facial tissue stiffness), detects hand-head contacts and is trained on a new dataset of facial performances. In the comprehensive experiments, *Decaf* demonstrates the highest 3D reconstruction (in terms of PVE) and plausibility metrics (in terms of F-score) among all compared methods. Especially significant are the advancement in terms of PVE compared to the most closely related previous method (roughly fourfold error reduction) and qualitative improvements in the estimated 3D geometry, which opens up many possibilities for downstream applications (*e.g.,* next-generation telepresence systems).

ACKNOWLEDGMENTS

We are very grateful to Weipeng Xu for his insightful discussions. All data captures were performed by MPII. The authors from MPII were supported by the ERC Consolidator Grant 4DRepLy (770784) and VIA research center. We acknowledge the support from Valeo.

REFERENCES

Jascha Achenbach, Robert Brylka, Thomas Gietzen, Katja zum Hebel, Elmar Schömer, Ralf Schulze, Mario Botsch, and Ulrich Schwanecke. 2018. A multilinear model for bidirectional craniofacial reconstruction. In *Proceedings of the Eurographics Workshop on Visual Computing for Biology and Medicine*. 67–76.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375* (2018).

Aljaz Bozic, Pablo Palafox, Michael Zollöfer, Angela Dai, Justus Thies, and Matthias Nießner. 2020. Neural Non-Rigid Tracking. (2020).

Adrian Bulat and Georgios Tzimiropoulos. 2017. How far are we from solving the 2D & 3D Face Alignment problem? (and a dataset of 230,000 3D facial landmarks). In *International Conference on Computer Vision (ICCV)*.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. 2021. Reconstructing hand-object interactions in the wild. In *International Conference on Computer Vision (ICCV)*.

Rishabh Dabral, Soshi Shimada, Arjun Jain, Christian Theobalt, and Vladislav Golyanik. 2021. Gravity-Aware Monocular 3D Human-Object Reconstruction. In *International Conference on Computer Vision (ICCV)*.

Radek Danecek, Michael J. Black, and Timo Bolkart. 2022. EMOCA: Emotion Driven Monocular Face Capture and Animation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 20311–20322.

Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. 2021a. Collaborative Regression of Expressive Bodies using Moderation. In *International Conference on 3D Vision (3DV)*. 792–804. https://doi.org/10.1109/3DV53792.2021.00088

Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021b. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (ToG)* 40, 4 (2021), 1–13.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. 2020. Three-dimensional reconstruction of human interactions. In *Computer Vision and Pattern Recognition (CVPR)*.

Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. 2021. Learning complex 3D human self-contact. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

David Fuentes-Jimenez, Daniel Pizarro, David Casillas-Perez, Toby Collins, and Adrien Bartoli. 2021. Texture-Generic Deep Shape-From-Template. *IEEE Access* 9 (2021), 75211–75230.

Pablo Garrido, Levi Valgaerts, Chenglei Wu, and Christian Theobalt. 2013. Reconstructing detailed dynamic face geometry from monocular video. *ACM Trans. Graph.* 32, 6 (2013), 158–1.

Pablo Garrido, Michael Zollhöfer, Chenglei Wu, Derek Bradley, Patrick Pérez, Thabo Beeler, and Christian Theobalt. 2016. Corrective 3D reconstruction of lips from monocular video. *ACM Trans. Graph.* 35, 6 (2016), 219–1.

Erik Gärtner, Mykhaylo Andriluka, Erwin Coumans, and Cristian Sminchisescu. 2022a. Differentiable dynamics for articulated 3d human motion reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*.

Erik Gärtner, Mykhaylo Andriluka, Hongyi Xu, and Cristian Sminchisescu. 2022b. Trajectory optimization for physics-based reconstruction of 3d human pose from monocular video. In *Computer Vision and Pattern Recognition (CVPR)*.

Vladislav Golyanik, Soshi Shimada, Kiran Varanasi, and Didier Stricker. 2018. Hdm-net: Monocular non-rigid 3d reconstruction with learned deformation model. In *Virtual Reality and Augmented Reality: 15th EuroVR International Conference, EuroVR 2018, London, UK, October 22–23, 2018, Proceedings 15*. Springer, 51–72.

Patrick Grady, Chengcheng Tang, Christopher D. Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C. Kemp. 2021. ContactOpt: Optimizing Contact to Improve Grasps. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiwen Guo, Feng Xu, Tao Yu, Xiaoyang Liu, Qionghai Dai, and Yebin Liu. 2017. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (ToG)* 36, 4 (2017), 1.

Marc Habermann, Weipeng Xu, Helge Rhodin, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2018. NRST: Non-rigid Surface Tracking from Monocular Video. In *German Conference on Pattern Recognition (GCPR)*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*.

Haoyu Hu, Xinyu Yi, Hao Zhang, Jun-Hai Yong, and Feng Xu. 2022. Physical Interaction: Reconstructing Hand-object Interactions with Physics. In *SIGGRAPH Asia 2022 Conference Papers*.

Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. 2022. Neural MoCon: Neural Motion Control for Physically Plausible Human Motion Capture. In *Computer Vision and Pattern Recognition (CVPR)*.

Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. 2015. Dynamic 3D avatar creation from hand-held video input. *ACM Transactions on Graphics (ToG)* 34, 4 (2015), 1–14.

Matthias Innmann, Michael Zollhöfer, Matthias Nießner, Christian Theobalt, and Marc Stamminger. 2016. Volumedeform: Real-time volumetric non-rigid reconstruction. In *International Conference on Computer Vision (ICCV)*.

Navami Kairanda, Edgar Tretschk, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. 2022. ϕ-SfT: Shape-from-Template with a Physics-based Deformation Model. In *Computer Vision and Pattern Recognition (CVPR)*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.

Yen Lee Angela Kwok, Jan Gralton, and Mary-Louise McLaws. 2015. Face touching: a frequent habit that has implications for hand hygiene. *American journal of infection control* 43, 2 (2015), 112–114.

Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. 2020. AvatarMe: Realistically Renderable 3D Facial Reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 760–769.

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 (TOG)* 36, 6 (2017), 194:1–194:17.

Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. 2022. MoCapDeform: Monocular 3D Human Motion Capture in Deformable Scenes. In *International Conference on 3D Vision (3DV)*.

Wenbin Lin, Chengwei Zheng, Jun-Hai Yong, and Feng Xu. 2022. Occlusionfusion: Occlusion-aware motion estimation for real-time dynamic 3d reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*.

Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. 2021. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Computer Vision and Pattern Recognition (CVPR)*.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for perceiving and processing reality. In *Workshop on Computer Vision for AR/VR at Computer Vision and Pattern Recognition (CVPRW)*.

Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. 2021. Dynamics-regulated kinematic policy for egocentric pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)* (2021).

Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. 2022. Embodied Scene-aware Human Pose Estimation. *Advances in Neural Information Processing Systems (NeurIPS)* (2022).

Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*.

Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. 2019. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG)* 38, 4 (2019).

Lea Müller, Ahmed A. A. Osman, Siyu Tang, Chun-Hao P. Huang, and Michael J. Black. 2021. On Self-Contact and Human Pose. In *Computer Vision and Pattern Recognition (CVPR)*.

Matthias Müller, Bruno Heidelberger, Marcus Hennix, and John Ratcliff. 2007. Position Based Dynamics. *J. Vis. Comun. Image Represent.* 18, 2 (apr 2007), 109–118.

Dat Tien Ngo, Sanghyuk Park, Anne Jorstad, Alberto Crivellaro, Chang D. Yoo, and Pascal Fua. 2015. Dense Image Registration and Deformable Surface Reconstruction in Presence of Occlusions and Minimal Texture. In *International Conference on Computer Vision (ICCV)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Antoine Petit, Stéphane Cotin, Vincenzo Lippiello, and Bruno Siciliano. 2018. Capturing deformations of interacting non-rigid objects using rgb-d data. In *International Conference on Intelligent Robots and Systems (IROS)*.

Pexels. 2023. Pexels. https://www.pexels.com/. Accessed: 2023-10-11.

Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. 2020. Contact and Human Dynamics from Monocular Video. In *European Conference on Computer Vision (ECCV)*.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics (TOG)* 36, 6 (Nov. 2017).

Shunsuke Saito, Tianye Li, and Hao Li. 2016. Real-time facial segmentation and performance capture from rgb input. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 244–261.

Mathieu Salzmann, Julien Pilet, Slobodan Ilic, and Pascal Fua. 2007. Surface Deformation Models for Nonrigid 3D Shape Recovery. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 29, 8 (2007), 1481–1487.

Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. Background Matting: The World is Your Green Screen. In *Computer Vision and Pattern Regognition (CVPR)*.

Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. 2022. HULC: 3D HUman Motion Capture with Pose Manifold SampLing and Dense Contact Guidance. In *European Conference on Computer Vision (ECCV)*.

Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Didier Stricker. 2019. Ismogan: Adversarial learning for monocular non-rigid 3d reconstruction. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, Patrick Pérez, and Christian Theobalt. 2021. Neural Monocular 3D Human Motion Capture with Physical Awareness. *ACM Transactions on Graphics (TOG)* 40, 4, Article 83 (aug 2021).

Soshi Shimada, Vladislav Golyanik, Weipeng Xu, and Christian Theobalt. 2020. PhysCap: Physically Plausible Monocular 3D Motion Capture in Real Time. *ACM Transactions on Graphics* 39, 6, Article 235 (dec 2020).

Miroslava Slavcheva, Maximilian Baust, Daniel Cremers, and Slobodan Ilic. 2017. Killingfusion: Non-rigid 3d reconstruction without correspondences. In *Computer Vision and Pattern Recognition (CVPR)*.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems (NeurIPS)* (2015).

Bugra Tekin, Federica Bogo, and Marc Pollefeys. 2019. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Computer Vision and Pattern Recognition (CVPR)*.

Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. 2017. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*.

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2387–2395.

Edith Tretschk, Navami Kairanda, Mallikarjun B R, Rishabh Dabral, Adam Kortylewski, Bernhard Egger, Marc Habermann, Pascal Fua, Christian Theobalt, and Vladislav Golyanik. 2023. State of the Art in Dense Monocular Non-Rigid 3D Reconstruction. *Computer Graphics Forum (EG STAR 2023)* (2023).

Aggeliki Tsoli and Antonis A Argyros. 2018. Joint 3D tracking of a deformable object in interaction with a hand. In *European Conference on Computer Vision (ECCV)*.

Jiayi Wang, Diogo Luvizon, Franziska Mueller, Florian Bernard, Adam Kortylewski, Dan Casas, and Christian Theobalt. 2022. HandFlow: Quantifying View-Dependent 3D Ambiguity in Two-Hand Reconstruction with Normalizing Flow. *Vision, Modeling, and Visualization* (2022).

Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. 2016. An anatomically-constrained local deformation model for monocular face capture. *ACM transactions on graphics (TOG)* 35, 4 (2016), 1–12.

Kevin Xie, Tingwu Wang, Umar Iqbal, Yunrong Guo, Sanja Fidler, and Florian Shkurti. 2021. Physics-based human motion estimation and synthesis from videos. In *International Conference on Computer Vision (ICCV)*.

Xinyu Yi, Yuxiao Zhou, Marc Habermann, Soshi Shimada, Vladislav Golyanik, Christian Theobalt, and Feng Xu. 2022. Physical Inertial Poser (PIP): Physics-aware Real-time Human Motion Tracking from Sparse Inertial Sensors. In *Computer Vision and Pattern Recognition (CVPR)*.

Alex Yu. 2023. Triangle mesh to signed-distance function (SDF). https://github.com/sxyu/sdf.

Rui Yu, Chris Russell, Neill DF Campbell, and Lourdes Agapito. 2015. Direct, dense, and deformable: Template-based non-rigid 3d reconstruction from rgb video. In *International Conference on Computer Vision (ICCV)*.

Ye Yuan, Shih-En Wei, Tomas Simon, Kris Kitani, and Jason Saragih. 2021. Simpoe: Simulated character control for 3d human pose estimation. In *Computer vision and pattern recognition (CVPR)*.

Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. 2021a. Interacting two-hand 3d pose and shape reconstruction from single color image. In *International Conference on Computer Vision (ICCV)*.

Hao Zhang, Zi-Hao Bo, Jun-Hai Yong, and Feng Xu. 2019. InteractionFusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. *ACM Transactions on Graphics (TOG)* 38, 4 (2019).

Hao Zhang, Yuxiao Zhou, Yifei Tian, Jun-Hai Yong, and Feng Xu. 2021b. Single depth view based real-time reconstruction of hand-object interactions. *ACM Transactions on Graphics (TOG)* 40, 3 (2021).

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5745–5753.

Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. 2014. Real-time non-rigid reconstruction using an RGB-D camera. *ACM Transactions on Graphics (ToG)* 33, 4 (2014).