DICE: END-TO-END DEFORMATION CAPTURE OF HAND-FACE INTERACTIONS FROM A SINGLE IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing 3D hand-face interactions with deformations from a single image is a challenging yet crucial task with broad applications in AR, VR, and gaming. The challenges stem from self-occlusions during single-view hand-face interactions, diverse spatial relationships between hands and face, complex deformations, and the ambiguity of the single-view setting. The previous state-of-the-art, Decaf, employs a global fitting optimization guided by contact and deformation estimation networks trained on studio-collected data with 3D annotations. However, Decaf suffers from a time-consuming optimization process and limited generalization capability due to its reliance on 3D annotations of hand-face interaction data. To address these issues, we present *DICE*, the first end-to-end method for Deformation-aware hand-face Interaction reCovEry from a single image. DICE estimates the poses of hands and faces, contacts, and deformations simultaneously using a Transformer-based architecture. It features disentangling the regression of local deformation fields and global mesh vertex locations into two network branches, enhancing deformation and contact estimation for precise and robust hand-face mesh recovery. To improve generalizability, we propose a weakly-supervised training approach that augments the training set using in-the-wild images *without* 3D ground-truth annotations, employing the depths of 2D keypoints estimated by off-the-shelf models and adversarial priors of poses for supervision. Our experiments demonstrate that DICE achieves state-of-the-art performance on a standard benchmark and in-the-wild data in terms of accuracy and physical plausibility. Additionally, our method operates at an interactive rate (20 fps) on an Nvidia 4090 GPU, whereas Decaf requires more than 15 seconds for a single image. Our code will be publicly available upon publication.

1 INTRODUCTION

Hand-face interaction is a common behavior observed up to 800 times per day across all ages and genders (Spille et al., 2021). Therefore, faithfully recovering hand-face interactions with plausible deformations is an important task given its wide applications in AR/VR (Pumarola et al., 2018; Hu et al., 2017; Wei et al., 2019), character animation (Qin et al., 2023; Zhao et al., 2024), and human behavior analysis (Liu et al., 2022; Guo et al., 2023; Mueller et al., 2019). Given the speed requirement of downstream applications like AR/VR, fast and accurate 3D reconstruction of hand-face interactions is highly desirable. However, several challenges make monocular hand-face deformation and interaction recovery particularly challenging: **1**) self-occlusions involved in hand-face interaction, **2**) the diversity of hand and face poses, contacts, and deformations, and **3**) ambiguity in the single-view setting. Most existing methods (Rempe et al., 2020; Muller et al., 2021) only reconstruct hand (Romero et al., 2022) and face (Li et al., 2017) meshes, unified as a whole-body model (Loper et al., 2023; Pavlakos et al., 2019), without capturing contacts and deformations. A seminal advance, Decaf (Shimada et al., 2023), recovers hand-face interactions while accounting for both deformations and contacts. However, Decaf requires time-consuming optimization, which takes more than 15 seconds per image, rendering it unsuitable for interactive applications. Moreover, Decaf's iterative fitting process depends heavily on accurate initial estimates of hand and face keypoints, as well as contact points on their surfaces, which could fail when significant occlusion is present in the image (See Fig. 8). Additionally, Decaf cannot scale up their training to fruitful hand-face interaction data in the wild,

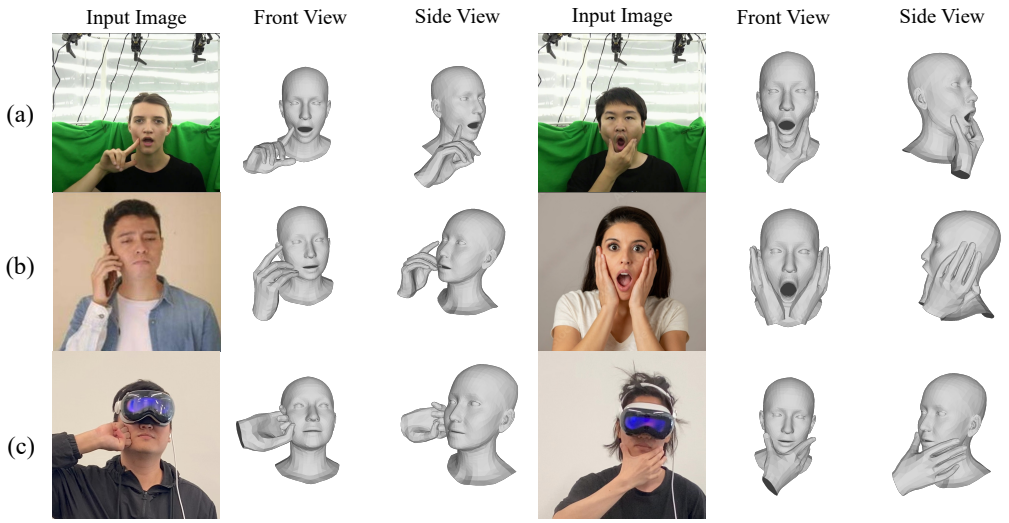


Figure 1: Our method is the first end-to-end approach that captures hand-face interaction and deformation from a monocular image. Results are from (a) Decaf's validation dataset, (b) in-the-wild images, and (c) VR use cases.

as they require 3D ground-truth annotations, such as contact labels and deformations that are not available from the in-the-wild data.

To tackle the issues above, we present *DICE*, the first end-to-end approach for Deformation-aware hand-face Interaction reCovEry from a monocular image. Our approach features three key designs: **1**) Our Transformer-based model leverages the attention mechanism to capture the relationships between the hand and face. **2**) Motivated by the global nature of pose and shape, as well as the local nature of the deformation field and contact probabilities–their invariance to global transformations of the hand and face–we propose disentangling the regression of global geometry and local interaction into two network branches. We evaluate this approach to enhance the estimation of deformations and contacts while ensuring accurate and robust recovery of hand and face meshes. **3**) Instead of directly regressing the hand and face parameters, we learn an intermediate non-parametric mesh representation. This representation is used to regress the pose and shape parameters of the hand and face using a neural inverse-kinematics network. Compared to directly regressing pose and shape parameters, which learns abstract parameters in a highly non-linear space, predicting vertex positions in Euclidean space and then applying inverse-kinematics improves the reconstruction accuracy (Li et al., 2021; 2023c;b). Combining all these contributions, our model achieves higher reconstruction accuracy than all previous regression- (Feng et al., 2021a; Li et al., 2017; Lin et al., 2021a) and optimization-based (Shimada et al., 2023; Lugaresi et al., 2019; Li et al., 2017) methods. Additionally, by utilizing the neural inverse-kinematics network, our approach benefits from an animatable parametric representation of the hand and face, which can be readily utilized in downstream applications.

Despite containing rich 3D annotations, the existing benchmark dataset (Shimada et al., 2023) collected in a studio is still limited in the diversity of hand motions, facial expressions, and appearances. Training only on such a dataset limits the model's ability to generalize to in-the-wild scenarios. To achieve robust and generalizable hand-face interaction and deformation recovery, we introduce a weak-supervision training pipeline that utilizes in-the-wild images without the reliance on 3D annotations. To achieve this, our key insight is to leverage additional prior knowledge, such as depth supervision alongside 2D keypoint supervision, compensating for the absence of ground truth contact and deformation annotations. We leverage the robust depth prior provided by a diffusion-based monocular depth estimation model (Ke et al., 2024), which provides essential geometric information for accurate mesh recovery and captures spatial relationships critical for contact state and deformation estimation. As the task becomes highly ill-posed for in-the-wild images, we further employ pose priors of the hand and face by introducing hand and face parameter discriminators that learn rich hand and face motion priors from additional datasets on hand or face separately (Pan et al., 2023a; Zimmermann et al., 2019). By incorporating a small set of real-world images alongside the Decaf dataset and leveraging our weak-supervision pipeline, we markedly enhance the accuracy and generalization capacity of our model.

As a result, our method achieves superior performance in terms of accuracy, physical plausibility, inference speed, and generalizability. It surpasses all previous methods in accuracy on both standard benchmarks and challenging in-the-wild images. Fig. 1 visualizes some results of our method. We conduct extensive experiments to validate our method. In summary, our contribution is three-fold:

- We propose DICE, the first end-to-end learning-based approach that accurately recovers hand-face interactions and deformations from a single image.
- We propose a novel weak-supervised training scheme with depth supervision on keypoints to augment the Decaf data distribution with a diverse real-world data distribution, significantly improving the generalization ability.
- DICE achieves superior reconstruction quality compared to baseline methods while running at an interactive rate (20fps).

2 RELATED WORK

Extensive efforts have been made to recover meshes from monocular images, including human bodies (Bogo et al., 2016; Moon & Lee, 2020; Li et al., 2021; Cai et al., 2024; Contributors; Xie et al., 2022; Wang & Daniilidis, 2023; Wang et al., 2023b; Lin et al., 2021b; Kanazawa et al., 2018; Cai et al., 2022; Zhang et al., 2021b; Feng et al., 2023; Li et al., 2022c; Wang et al., 2023a; Dou et al., 2023b; Cho et al., 2022; Huang et al., 2022b; Lin et al., 2021a), hands (Rong et al., 2021; Moon et al., 2020; 2024; Moon, 2023; Oh et al., 2023; Park et al., 2022; Yang et al., 2021; 2022b; Li et al., 2023d; Yu et al., 2023), and faces (Feng et al., 2021b; 2018; Wood et al., 2022; Daněček et al., 2022; Zielonka et al., 2022; Chai et al., 2023; Zhang et al., 2023c; Otto et al., 2023; He et al., 2023; Chatziagapi & Samaras, 2023; Kumar et al., 2023; Li et al., 2023a). This also includes recovering the surrounding environments (Clever et al., 2022; Huang et al., 2022a; Hassan et al., 2019; 2021; Zhang et al., 2020b; Li et al., 2022b; Zhang et al., 2021c; Shimada et al., 2022; Luo et al., 2022; Weng & Yeung, 2021) and interacting objects (Yang et al., 2022a; Zhang et al., 2020a; Pham et al., 2017; Tsoli & Argyros, 2018; Hampali et al., 2020; Tekin et al., 2019; Zhang et al., 2020a; Grady et al., 2021; Pokhariya et al., 2023; Hasson et al., 2019; Ye et al., 2022; Chen et al., 2023; 2021; Liu et al., 2021; Corona et al., 2020) while reconstructing the mesh. The acquired versatile behaviors play a crucial role in various applications, including motion generation (Tevet et al., 2022; Peng et al., 2022; Pan et al., 2023b; Guo et al., 2022; Wang et al., 2022a; Xu et al., 2023; 2024; Lin et al., 2024; Zhou et al., 2023; Wan et al., 2023a; Peng et al., 2021; Dou et al., 2023a; Wan et al., 2023b), augmented reality (AR), virtual reality (VR), and human behavior analysis (Zhang et al., 2023a; Yang et al., 2024; Zhang et al., 2024; 2023b; Guo et al., 2023; Liu et al., 2022). In the following, we mainly review the related works on hand, face and full-body mesh recovery.

3D Interacting Hands Recovery. Recent advancements have markedly enhanced the capture and recovery of 3D hand interactions. Early studies have achieved reconstruction of 3D hand-hand interactions utilizing a fitting framework, employing resources such as RGBD sequences (Oikonomidis et al., 2012), hand segmentation maps (Mueller et al., 2019), and dense matching maps (Wang et al., 2020). The introduction of large-scale datasets for interacting hands (Moon et al., 2020; 2024) has motivated the development of regression-based approaches. Notably, these include regressing 3D interacting hand directly from monocular RGB images (Rong et al., 2021; Moon, 2023; Zhang et al., 2021a; Li et al., 2022a; Zuo et al., 2023). Additionally, research has extended to recovering interactions between hands and various objects in the environment, including rigid (Cao et al., 2021; Grady et al., 2021; Liu et al., 2021; Tekin et al., 2019; Fan et al., 2024; Ye et al., 2023b;a), articulated (Fan et al., 2023), and deformable (Tretschk et al., 2023) objects. Following Shimada et al. (2023), our work distinguishes itself by introducing hand interactions with a deformable face, characterized by its non-uniform stiffness—a significant difference from conventional deformable models. This innovation presents unique challenges in accurately modeling interactions.

3D Human Face Recovery. Research in human face recovery encompasses both optimization-based (Aldrian & Smith, 2012; Thies et al., 2016) and regression-based (Feng et al., 2018; Sanyal et al., 2019) methodologies. Beyond mere geometry reconstruction, recent approaches have evolved to incorporate training networks with the integration of differentiable renderers (Feng et al., 2021b; Zielonka et al., 2022; Zheng et al., 2022; Wang et al., 2022b; Cho et al., 2022). These methods estimate variables such as lighting, albedo, and normals to generate facial images and compare them with the monocular input. However, a significant limitation in much of the existing literature is the neglect of the face's deformable nature and hand-face interactions. Decaf (Shimada et al., 2023) represents a pivotal development in this area, attempting to model the complex mimicry of musculature and the underlying skull anatomy through optimization techniques. In contrast, our work introduces a regression-based, end-to-end method for efficient problem-solving, setting a new benchmark in the field.

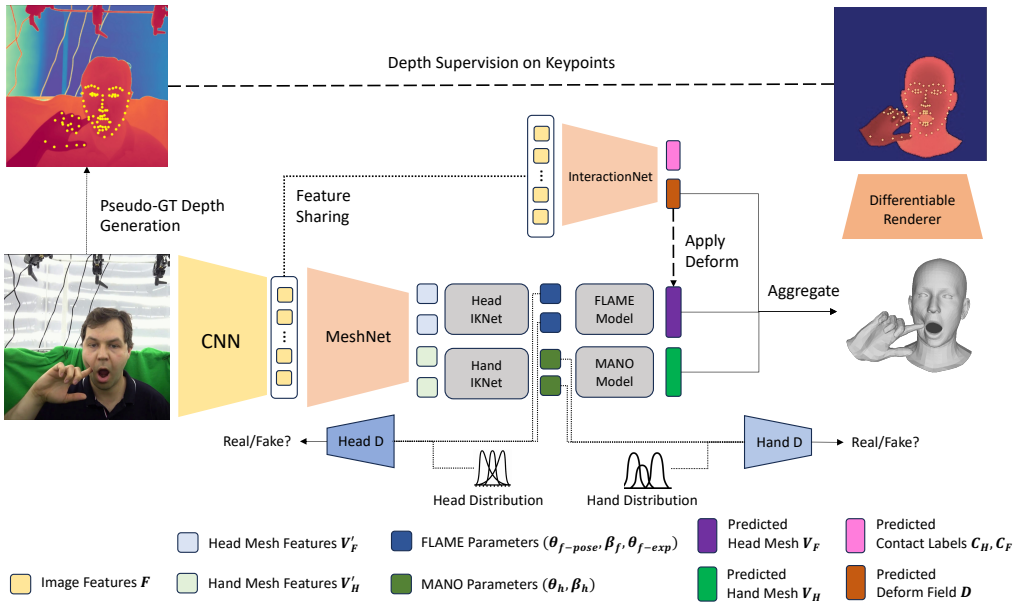


Figure 2: **Overview of the proposed DICE framework**. The input image is first fed to a CNN to extract a feature map, which is then passed to the Transformer-based encoders for mesh and interaction, *i.e.*, MeshNet and InteractionNet. MeshNet extracts hand and face mesh features, which are then used by the Inverse Kinematics models (IKNets) to predict pose and shape parameters that drive FLAME (Li et al., 2017) and MANO (Romero et al., 2022) models. InteractionNet predicts per-vertex hand-face contact probabilities and face deformation fields from the feature map, where the latter is applied to the face mesh output by the FLAME model. To improve the generalization capability, we introduce a weakly-supervised training scheme using off-the-shelf 2D keypoint detection models (Lugaresi et al., 2019; Bulat & Tzimiropoulos, 2017) and depth estimation models (Ke et al., 2024) to provide depth supervision on keypoints. In addition, we use face and hand discriminators to constrain the distribution of parameters regressed by IKNets.

3D Full-Body Recovery. The task of monocular human pose and shape estimation involves reconstructing a 3D human body from a single image. Optimization-based approaches (Bogo et al., 2016; Pavlakos et al., 2019; Shi et al., 2023; Rempe et al., 2021) employ the SMPL model (Loper et al., 2023), fitting it to 2D keypoints detected within the image. Conversely, regression-based methods (Li et al., 2021; Lassner et al., 2017; Kocabas et al., 2021; Kanazawa et al., 2018; Feng et al., 2021a; Fang et al., 2021; Lin et al., 2023; Cai et al., 2024; Feng et al., 2023) leverage deep neural networks to directly infer the pose and shape parameters of the SMPL model. Hybrid methods (Kolotouros et al., 2019a) integrate both optimization and regression techniques, enhancing 3D model supervision. Distinct from these approaches, we follow parametric methods (Li et al., 2021; Cai et al., 2024; Kanazawa et al., 2018; Bogo et al., 2016) due to its flexibility for animation purposes. Unlike most research in this domain, which primarily concentrates on the main body with only rough estimations of hands and face, our methodology uniquely accounts for detailed interactions between these components.

3 METHOD

Problem Formulation. Following Decaf (Shimada et al., 2023), we adopt the FLAME (Li et al., 2017) and MANO (Romero et al., 2022) parametric models for hand and face. Given a single RGB image $\mathbf{I} \in \mathbb{R}^{224\times224\times3}$, the objective of this task is to reconstruct the vertices of a hand mesh $\mathbf{V}_H \in \mathbb{R}^{778\times3}$ and a face mesh $\mathbf{V}_F \in \mathbb{R}^{5023\times3}$, along with capturing the face deformation vectors $\mathbf{D} \in \mathbb{R}^{5023\times3}$ resulting from hand-face interaction and its non-rigid nature. Additionally, we estimate per-vertex contact probabilities of hand $\mathbf{C}_H \in \mathbb{R}^{778}$ and face $\mathbf{C}_F \in \mathbb{R}^{5023}$.

3.1 TRANSFORMER-BASED HAND-FACE INTERACTION RECOVERY

Our model incorporates a two-branch Transformer architecture and integrates inverse-kinematic models, specifically, MeshNet, InteractionNet, and IKNets. A differentiable renderer (Ravi et al., 2020) is used to compute depth maps from the predicted mesh for depth supervision, while the hand and face discriminators are used as priors for constraining the hand and face poses; See Fig. 2 for an overview.

Given a monocular RGB image $\mathbf{I}$, we use a pretrained HRNet-W64 (Sun et al., 2019) backbone to extract a feature map $\boldsymbol{X}_{\mathrm{I}} \in \mathbb{R}^{H \times W \times C}$. Following Lin et al. (2021a;b), we flatten the image feature maps and upsample the $H \times W$ feature maps to N feature maps, corresponding to each keypoint and downsampled vertex of both hand and face. The feature maps $\mathbf{F}' \in \mathbb{R}^{N \times C}$ are then concatenated with the downsampled hand and face vertex and keypoint coordinates of dimension $N \times 3$, with the pose set to the mean pose, serving as positional encodings. This results in the final feature map $\mathbf{F} \in \mathbb{R}^{N \times (C+3)}$. To model the vertex-vertex interactions, we mask the feature maps $\mathbf{F}$ for a randomly selected subset of vertices.

Once the feature map $\mathbf{F}$ is obtained, it is fed into MeshNet and InteractionNet, which handle the regression of mesh vertices and the deformation field separately. This decomposition is motivated by their semantic differences: mesh contains global features, whereas deformation vectors and contact states are localized features, *i.e.*, invariant to the global transformations of the hand and face. Thus, MeshNet takes the feature map $\mathbf{F}$ as input and regresses the unrefined vertex positions of hand $\mathbf{V}'_H$ and face $\mathbf{V}'_F$. InteractionNet, on the other hand, predicts the 3D deformation field $\mathbf{D}$ for each face vertex, along with the contact labels for each hand and face vertex, $\mathbf{C}_H$ and $\mathbf{C}_F$, respectively. Note the contacts and deformations are regressed within the same encoder to model their causal relationship: the contacts cause the deformations. We validate our design in Sec. 4.4.

Next, instead of directly using the unrefined hand and face vertices $\mathbf{V}'_H$ and $\mathbf{V}'_F$, our method takes these vertices as input to regress the pose and shape of their respective parametric models (Li et al., 2017; Romero et al., 2022). This is achieved by a neural inverse kinematics model, named IKNet, following Kolotouros et al. (2019b). The IKNet takes the unrefined hand and face vertices $\mathbf{V}'_H$ and $\mathbf{V}'_F$ as inputs and predicts their pose, shape, and expression parameters $(\boldsymbol{\theta}_{\mathrm{h}}, \boldsymbol{\beta}_{\mathrm{h}})$ for hand, $(\boldsymbol{\theta}_{\text{f-pose}}, \boldsymbol{\beta}_{\mathrm{f}}, \boldsymbol{\theta}_{\text{f-exp}})$ for face, along with the root position and orientation for hand $(\boldsymbol{t}_h, \boldsymbol{r}_h)$ and face $(\boldsymbol{t}_f, \boldsymbol{r}_f)$, respectively. Afterward, we use the predicted parameters to first obtain the hand vertices $\mathbf{V}_H$ and undeformed face vertices $\mathbf{V}^*_F$. Then, we apply the deformation $\mathbf{D}$ predicted by the InteractionNet on $\mathbf{V}^*_F$ to get the final deformed face $\mathbf{V}_F$. Utilizing parametric forward-kinematics and neural inverse-kinematics models offer several advantages: first, it enables readily animatable meshes for downstream applications; second, compared to non-parametric regression methods, where meshes typically contain artifacts such as spikes (Lin et al., 2021a; Cho et al., 2022; Lin et al., 2021b), this approach significantly improves mesh quality; third, the compact parameter space allows for a more effective discriminator, which will be discussed in the following section.

3.2 Weakly-Supervised Training Scheme

Although the aforementioned benchmark, Decaf (Shimada et al., 2023), accurately captures hand, face, self-contact, and deformations, it consists of only eight subjects and is recorded in a green-screen studio. Thus, training a model only with the Decaf dataset limits its generalization capability to in-the-wild images that exhibit far more complex and diverse human identities, hand poses, and face poses.

To further enhance the generalization capability, we train our model with 500 diverse in-the-wild images of hand-face interaction collected from the internet *without* the reliance on the 3D ground truth annotations. First, we use 2D hand and face keypoints detected by Lugaresi et al. (2019) and Bulat & Tzimiropoulos (2017) as pseudo-ground-truth. Then, we use Marigold (Ke et al., 2024), a diffusion-based monocular depth estimator pre-trained on a large amount of images to generate 2D affine-invariant depth maps for depth supervision (see Eq. 4). The depth supervision provides a strong depth prior, which guides the spatial relationship between hand and face meshes, promoting accurate modeling of hand-face interaction. We first use a differentiable rasterizer (Ravi et al., 2020) to compute a depth map from the predicted hand and face meshes. We use a depth loss to measure the difference between the depths of the hand and face keypoints and their corresponding points on the predicted depth map, providing supervision. This keypoint-to-keypoint correspondence enables accurate depth supervision even when the rendered hand/face mesh and the ground-truth meshes are misaligned. Moreover, we train adversarial priors on the hand and face parameter space on multiple hand and face motion datasets: the face-only RenderMe-360 (Pan et al., 2023a), the hand-only FreiHand (Zimmermann et al., 2019), and Decaf (Shimada et al., 2023). This ensures the plausibility of generated face and hand poses and shapes while allowing for flexible poses and shapes beyond the Decaf data distribution to handle in-the-wild cases. The overall weak-supervision pipeline significantly enhances our model's generalization capability and robustness, which we investigate in Sec. 4.4.

3.3 Loss Functions

Mesh losses $\mathcal{L}_{\text{mesh}}$: For richly annotated Decaf dataset (Shimada et al., 2023), we employ an L_1 loss for 3D keypoints, 3D vertices, and 2D reprojected keypoints, comparing them against their respective ground-truths, following common practice in human and hand mesh recovery (Lin et al., 2021a; Cho et al., 2022; Dou et al., 2023b). We further apply an L_1 loss $\mathcal{L}_{\text{params}}$ on the estimated hand and face pose, shape, and facial expression against the ground-truth parameters. For in-the-wild data, only the 2D reprojected keypoints are supervised, as they are the only type with corresponding ground truth.

Interaction losses $\mathcal{L}_{\text{interaction}}$: Similar to Shimada et al. (2023), we impose Chamfer Distance losses to promote touch for predicted contact vertices and discourage collision. We also introduce a binary cross-entropy loss to supervise contact labels and a deform loss with adaptive weighting mechanism to supervise deform vectors. For in-the-wild data, we also impose touch and collision losses since they do not require annotations.

Adversarial loss $\mathcal{L}_{\text{adv}}$ are applied to the predicted hand and face parameters for in-the-wild data to constrain their parameter space, and for Decaf data to facilitate the training of the discriminators. The adversarial loss is given by:

$$\mathcal{L}_{\text{adv}}(E) = \mathbb{E}_{\boldsymbol{\theta}_f \sim p_E}\left[\log\left(1 - D_F(E(I))\right)\right] + \mathbb{E}_{\boldsymbol{\theta}_h \sim p_E}\left[\log\left(1 - D_H(E(I))\right)\right]. \tag{1}$$

The losses for the hand and face discriminators are given by:

$$\mathcal{L}_{\text{adv}}(D_F) = -\left(\mathbb{E}_{\boldsymbol{\theta}_f \sim p_E}\left[\log\left(1 - D_F(E(I))\right)\right] + \mathbb{E}_{\boldsymbol{\theta}_f \sim p_{\text{data}}}\left[\log\left(D_F(\boldsymbol{\theta}_f)\right)\right]\right), \tag{2}$$

and

$$\mathcal{L}_{\text{adv}}(D_H) = -\left(\mathbb{E}_{\boldsymbol{\theta}_H \sim p_E}\left[\log\left(1 - D_H(E(I))\right)\right] + \mathbb{E}_{\boldsymbol{\theta}_H \sim p_{\text{data}}}\left[\log\left(D_H(\boldsymbol{\theta}_h)\right)\right]\right), \tag{3}$$

where E jointly denotes the image backbone, the mesh encoder and the parameter regressor, p_E denotes the output distribution of E, p_{data} denotes the data distribution of the motion datasets, $\boldsymbol{\theta}_f = (\boldsymbol{\theta}_{\text{f-pose}}, \boldsymbol{\beta}_{\text{f}}, \boldsymbol{\theta}_{\text{f-exp}})$, $\boldsymbol{\theta}_H = (\boldsymbol{\theta}_{\text{h}}, \boldsymbol{\beta}_{\text{h}})$.

Depth loss $\mathcal{L}_{\text{depth}}$: To provide pseudo-3D hand and face keypoints supervision for in-the-wild data, we use a modified SILog Loss (Eigen et al., 2014), an affine-invariant depth loss as our depth supervision $\mathcal{L}_{\text{depth}}$. Formally, let $\hat{K}_D$ denote the pseudo-ground-truth affine-invariant depth of the face and hand keypoints, and K_D denote the rendered depth for the keypoints,

$$\mathcal{L}_{\text{depth}} = \left[\mathbf{Var}\left(\log(K_D + \varepsilon) - \log(\hat{K}_D + \varepsilon)\right)\right]^{1/2}, \tag{4}$$

where $\mathbf{Var}$ is the standard variance operator and $\varepsilon = 10^{-7}$.

Overall, our loss for the mesh and interaction networks is formulated by

$$\mathcal{L} = \lambda_{\text{mesh}}\mathcal{L}_{\text{mesh}} + \lambda_{\text{interaction}}\mathcal{L}_{\text{interaction}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}, \tag{5}$$

where $\lambda_{\text{mesh}} = 12.5, \lambda_{\text{interaction}} = 5, \lambda_{\text{depth}} = 2.5, \lambda_{\text{adv}} = 1$ for all the experiments in the paper; See more details in Appendix C.

4 Experimental Results

4.1 Datasets and Metrics

Datasets We employ Decaf (Shimada et al., 2023) for reconstructing 3D face and hand interactions with deformations, along with the in-the-wild dataset we collected containing 500 images. We use the shape, pose, and expression data of hands and faces from Decaf (Shimada et al., 2023), RenderMe-360 (Pan et al., 2023a), and FreiHand (Zimmermann et al., 2019) for training the adversarial priors. We use the training set of the aforementioned datasets for network training. We use the official split from Decaf to separate the training and testing sets, and select a few in-the-wild images for the test set to perform qualitative visualizations.

Metrics We adopt commonly-used metrics for mesh recovery accuracy following Kanazawa et al. (2018); Lin et al. (2021a); Dou et al. (2023b); Cho et al. (2022): **1**) *Mean Per-Joint Position Error* (MPJPE): the average Euclidean distance between predicted keypoints and ground-truth keypoints. **2**) *PAMPJPE*: MPJPE after Procrustes Analysis (PA) alignment. **3**) *Per Vertex Error*: per vertex

Table 1: Comparison of hand-face interaction and deformation recovery on Decaf.

Methods	Type	3D Reconstruction Error			Physics Plausibility Metrics				Running Time (per image; s)↓
		PVE‡↓	MPJPE↓	PAMPJPE↓	Col. Dist. ↓	Non. Col. ↑	Touchness ↑	**F-Score** ↑	
Comparison between DICE and optimization-based methods									
Decaf (Shimada et al., 2023)	O	9.65	–	–	1.03	83.6	96.6	**89.6**	19.59
Benchmark (Lugaresi et al., 2019; Li et al., 2017)	O	17.7	–	–	19.3	64.2	73.2	68.4	16.40
PIXIE (hand+face) (Feng et al., 2021a)	O	26.3	–	–	7.04	75.9	75.1	75.5	–
DICE (Ours)	R	**8.32**	**9.95**	**7.27**	**0.16**	66.6	79.9	72.7	**0.088**
Comparison between DICE and regression-based methods									
PIXIE (whole-body) (Feng et al., 2021a)	R	39.7	–	–	0.11	97.1	51.8	67.6	**0.070**
PIXIE-R (Feng et al., 2021a)	R	11.0	22.0	21.2	0.27	62.6	83.0	72.0	**0.070**
METRO* (hand+face) (Lin et al., 2021a)	R	11.8	15.4	11.9	**0.08**	80.7	54.8	65.2	0.103
FastMETRO* (single-target) (Cho et al., 2022)	R	9.27	11.8	9.41	0.09	82.2	55.5	66.2	0.110
DICE (Ours)	R	**8.32**	**9.95**	**7.27**	0.16	66.6	79.9	**72.7**	0.088

* parametric version. O and R denote optimization-based and regression-based methods, respectively. ‡ calculated after translating the center of the head to the origin. **bold** denotes the best result in a comparison group. Note our method operates at an interactive rate (20 fps; 0.049s per image) on an Nvidia 4090 GPU. Here we report the runtime performance on an A6000 GPU for a fair comparison.

error (PVE) with translation. Following Decaf (Shimada et al., 2023), we use the following metrics to measure the plausibility: **4**) *Collision Distance* (Col. Dist.): the average collision distances over vertices and frames; **5**) *Non-Collision Ratio* (Non. Col.): the proportion of frames without hand-face collisions; **6**) *Touchness Ratio* (Touchness): the ratio of hand-face contacts among ground truth contacting frames; **7**) *F-Score*: the harmonic mean of *Non-Collision Ratio* and *Touchness Ratio*. Note that F-Score measures Touchness and Non-Collision Ratio as a whole, which is a metric of overall physical plausibility, whereas Non-Collision Ratio or Touchness are meaningless when considered individually.

4.2 Implementation Details

We train MeshNet, InteractionNet, and IKNet, along with the face and hand discriminators using AdamW (Loshchilov, 2017) optimizers, each with a learning rate of 6×10^{-4}, and a learning rate decay of 1×10^{-4}. The generator and discriminator networks are optimized in an alternating manner. Our batch size is set to 16 during the training stage. The training takes 40 epochs, totalling 48 hours. The model is trained and evaluated on 8 Nvidia A6000 GPUs with an AMD 128-core CPU. Inference times are calculated on a single Nvidia A6000 GPU.

4.3 Performance on Hand-Face Interaction and Deformation Recovery

We compare our method with the following: **1**) **Benchmark**: the baseline (Lugaresi et al., 2019; Li et al., 2017) introduced in Decaf (Shimada et al., 2023); **2**) **Decaf** (Shimada et al., 2023): an optimization-based method for hand-face interaction and deformation recovery. **3**) **PIXIE (whole-body)** (Feng et al., 2021a): a representative model for full-body recovery, including the hand and face, introduced in Decaf. **4**) **PIXIE (hand+face)** (Feng et al., 2021a): a optimization-based variant of PIXIE, introduced in Decaf. For regression-based methods, as we are dealing with a relatively new task, there are few readily available baselines. To facilitate comparison, we adapt the following regression-based models from related tasks: **5**) **METRO** (Lin et al., 2021a): A representative work in human body/hand mesh recovery. We adapt METRO to predict both hand and face meshes, adding extra output heads to predict contact and deformation. **6**) **PIXIE-R** (Feng et al., 2021a): Adapted PIXIE, using the same backbone and hand and face branches but trained with losses from DICE. **7**) **FastMETRO (single-target)** (Cho et al., 2022): Another representative work in human and hand mesh recovery. We adapt two independent FastMETROs, one for estimating hand mesh vertices and contact, and the other for estimating face mesh, deformation, and contact. Here, the word single-target means each FastMETRO considers hand and face individually, with no information exchange. This model is trained using the same hyperparameter, loss, and optimizer as DICE, on the Decaf (Shimada et al., 2023) dataset.

4.3.1 Quantitative Evaluations

Reconstruction Accuracy In Tab. 1, our method surpasses all baseline methods in terms of reconstruction accuracy, achieving a 7.5% reduction in per-vertex error compared to the current state-of-the-art, Decaf. Note that our method is regression-based and allows inference at an interactive rate, while Decaf (Shimada et al., 2023) uses a cumbersome test-time optimization process, taking more than 200x more time per image. Decaf also requires using temporal information in successive frames, while our method only uses a single frame. Our method shows a 30% reduction in reconstruction error

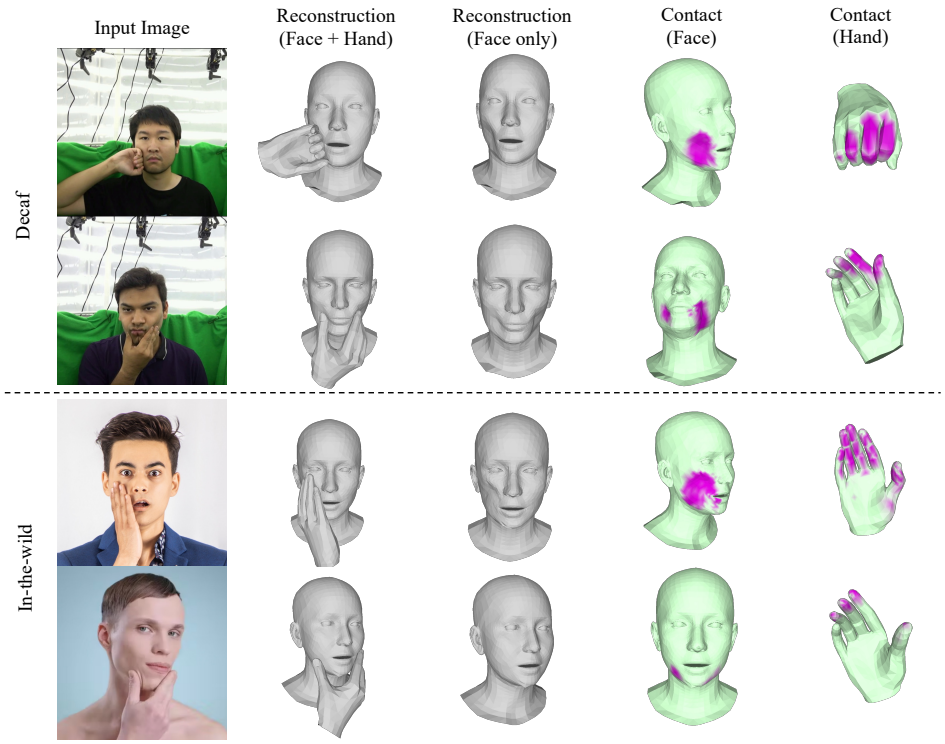


Figure 3: Qualitative results of hand-face interaction, deformation, and contact recovery by DICE on Decaf and in-the-wild images. In contact visualizations, a deeper color indicates a higher contact probability.

Table 2: Comparison of hand-face contact estimation on Decaf.

Method	**F-score** ↑	Precision ↑	Recall ↑	Accuracy↑
Decaf (face) (Shimada et al., 2023)	0.57	0.69	0.49	0.99
Decaf (hand) (Shimada et al., 2023)	0.47	0.62	0.39	0.98
DICE (face)	**0.61**	0.64	0.57	1.00
DICE (hand)	**0.50**	0.55	0.45	0.98

compared to the modified METRO baseline, and up to 79% reduction compared to other end-to-end baselines. Notably, our method achieves a 27% MPVE reduction compared to the PIXIE-R baseline which uses the same mesh and interaction losses as our method, demonstrating the superiority of our network design and weak-supervised training scheme. Our method is also more accurate than another end-to-end baseline, FastMETRO.

Plausibility In terms of overall physical plausibility (F-Score), our method is the best among all regression-based methods: PIXIE (whole-body), PIXIE-R, METRO, and FastMETRO. On the other hand, while some optimization-based methods (Decaf and PIXIE (hand+face) have higher overall plausibility (F-Score) compared to DICE, this is due to their test-time optimization, which iteratively adjusts the relative positioning of hand and face. Thus, they are much more computationally intensive than our regression-based method. With a highly efficient end-to-end inference scheme, DICE still outperformed an optimization-based method (Benchmark) on F-Score.

Contact Estimation The contact estimation metrics (accuracy, precision, recall) are calculated by the predicted per-vertex contact probabilities against the respective 0-1 contact ground truths. In Tab. 2, DICE achieves superior contact estimation performance on the Decaf dataset, surpassing previous work (Shimada et al., 2023) in F-Score for both face and hand contacts. Here F-score provides a comprehensive measure of both precision and recall ratio combined. These two metrics involve a trade-off: focusing solely on precision may lead to a decrease in recall, and vice versa. Balancing this trade-off, the F-score offers a more meaningful evaluation of contact estimation.

4.3.2 Qualitative Evaluations

As discussed in Sec. 3.2, the Decaf (Shimada et al., 2023) dataset is collected in an indoor environment with a green screen, which doesn't reflect the complex environment where real-world hand-face interactions occur. Therefore, a model only trained with the Decaf dataset might have generalization issues when tested on in-the-wild data. Fig. 4 supports this claim by demonstrating our model's superior generalization performance on in-the-wild data with unseen identity and pose. On the

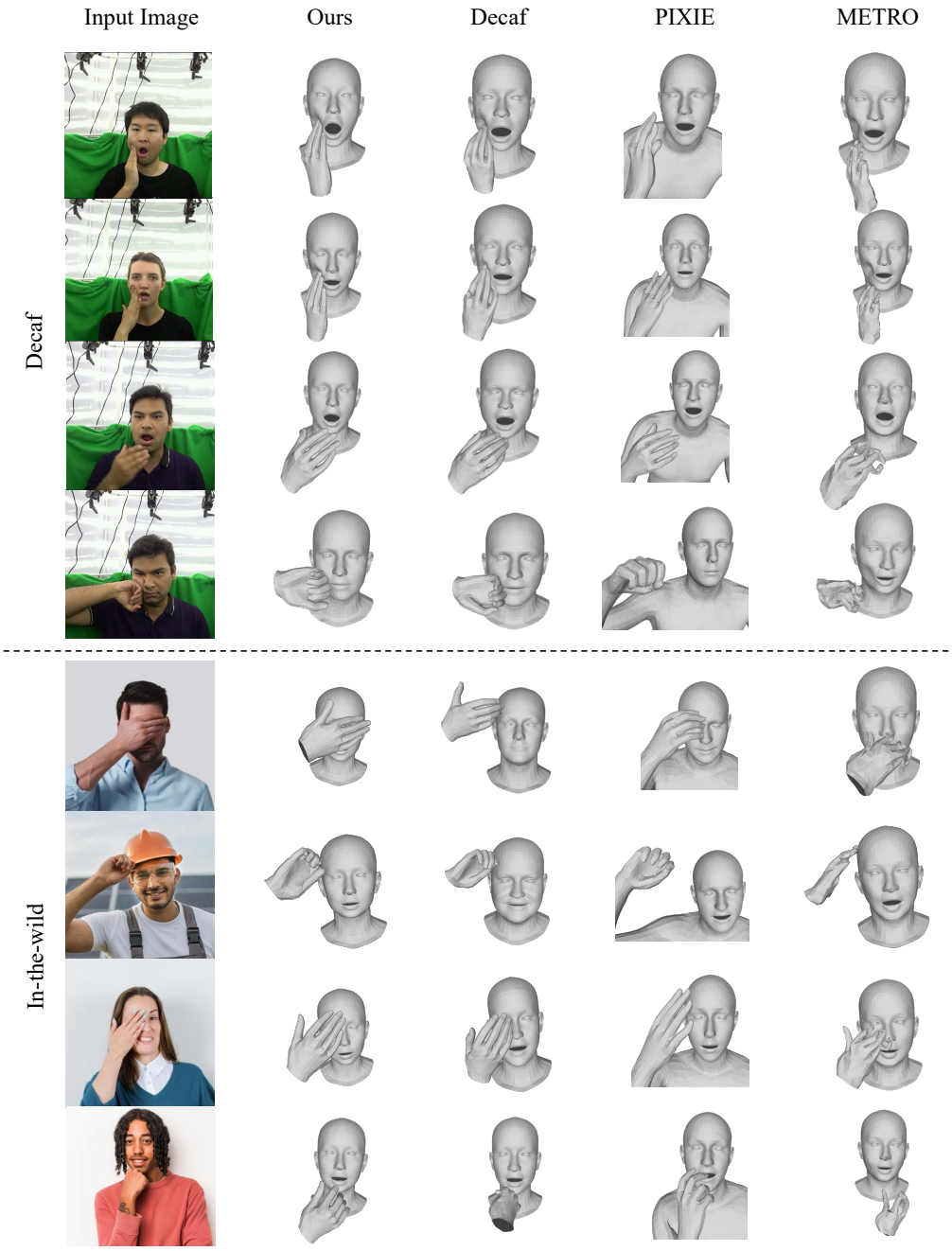


Figure 4: Qualitative comparison of DICE, Decaf (Shimada et al., 2023), PIXIE (Feng et al., 2021a) (whole-body version), METRO* (Lin et al., 2021b) on Decaf validation set and in-the-wild images. Our method achieves superior reconstruction accuracy and plausibility in the Decaf (Shimada et al., 2023) dataset, especially generalizing well to difficult in-the-wild actions unseen in Decaf compared to all baselines.

other hand, Decaf's reconstruction suffers from self-collision and incorrect hand-face relationship. PIXIE and METRO reconstruct inaccurate hand poses and often demonstrates implausible non-touching artefacts. As shown in Fig. 3, our method faithfully reconstructs hand-face interaction and deformation and accurately labels the contact areas.

4.4 Ablation Study

Network Design In Tab. 3, adopting the two-branch architecture, which separates deformation and interaction estimation from mesh vertices regression, improves both accuracy and plausibility.

In-the-wild data As shown in Tab. 3, adding weak-supervision training and in-the-wild data for DICE training improves all reconstruction error metrics (PVE*, MPJPE, PAMPJPE) while maintaining a high plausibility (F-Score). We deem that the slight decrease in F-Score could mainly be attributed to

Table 3: Comparison of hand-face interaction and deformation recovery on Decaf. **Bold** denotes the best result.

Methods	PVE*↓	MPJPE↓	PAMPJPE↓	F-Score ↑
DICE (single branch)	9.29	11.6	8.51	69.3
DICE (w.o. in-the-wild data)	8.93	11.0	7.50	**73.3**
DICE (w.o. supervision on $\mathbf{V}'_F$, $\mathbf{V}'_H$)	12.2	14.4	11.1	70.7
DICE (w.o. $\mathcal{L}_{\text{depth}}$)	15.6	19.5	13.7	64.2
DICE (w.o. $\mathcal{L}_{\text{params}}$)	10.3	12.8	10.4	64.7
DICE (w.o. $\mathcal{L}_{\text{adv}}$)	11.1	14.2	10.4	69.8
DICE (Full)	**8.32**	**9.95**	**7.27**	72.7

the difference in distribution between the studio-collected Decaf (Shimada et al., 2023) and in-the-wild data. This is because the limited pose and identity distribution of the Decaf training dataset may cause the model to overfit, and the inclusion of in-the-wild images out of the Decaf data distribution effectively improves the generalization capability of DICE.

Unrefined Features Supervision Regressing the unrefined head and hand mesh features $\mathbf{V}'_F$, $\mathbf{V}'_H$ and then perform inverse kinematics to regress the parametric mesh improves plausibility and accuracy, compared to directly estimating the face and hand parameters.

Depth Supervision Although depth supervision is only applied to in-the-wild data, as shown in Tab. 3, removing it also significantly degrades performance on the Decaf validation set. Without depth loss, wrong predictions in depth are not penalized for in-the-wild data, introducing noise in the training process, and resulting in erroneous predictions on the Decaf dataset. As shown in Appendix Fig. 7, the absence of depth supervision introduces ambiguity in the z-direction, resulting in artifacts such as self-collision.

Parameter Supervision Supervising parameters directly, in addition to the indirect supervision of parameters by the mesh losses, improves both plausibility and accuracy. This is because direct parameter supervision eliminates ambiguity, preventing the network from converging to alternative parameter combinations that produce incorrect meshes that appear geometrically similar, *i.e.*, with small vertex loss, to the target but are incorrect in their underlying structure, such as pose or shape.

Adversarial Prior The adversarial prior incorporates diverse but realistic pose and shape distribution beyond Decaf (Shimada et al., 2023), ensuring the reality of regressed mesh while allowing for generalization. As shown in Tab. 3, introducing adversarial supervision improves the accuracy and physical plausibility.

4.5 Limitations and Future Works

While our method achieves SotA accuracy on the Decaf (Shimada et al., 2023) dataset and generalizes well to unseen scenes and in-the-wild cases, it still encounters failure cases when the hand-pose interactions are extremely challenging and have severe occlusions (see Appendix D.2). Moreover, despite our method effectively recovering hand and face meshes with visually plausible face deformations, there remains room for improvement in deformation accuracy and physical plausibility. Hand deformations could also be considered in future work for more realistic reconstructions. In the future, physics-based simulation (Hu et al., 2018; Li et al., 2020; Hu et al., 2019; Han et al., 2019; Lin et al., 2022; Huang et al., 2024) can be used as a stronger prior, producing more physically accurate estimations. In this paper, although we found using 500 in-the-wild images significantly improves the model's generalization ability, scaling up to a larger amount of in-the-wild data, on the order of millions or billions, would further enhance performance, which we will study in future work.

5 Conclusion

In this work, we present DICE, the first end-to-end approach for reconstructing 3D hand and face interaction with deformation from monocular images. Our approach features a two-branch transformer structure, MeshNet, and InteractionNet, to model local deform field and global mesh geometry. An inverse-kinematic model, IKNet, is used to output the animatable parametric hand and face meshes. We also proposed a novel weak-supervision training pipeline, using a small amount of in-the-wild images and supervising with a depth prior and an adversarial loss to provide pose priors. Benefitting from our network design and training scheme, DICE demonstrates state-of-the-art accuracy and plausibility, compared with all previous methods. Meanwhile, our method achieves a fast inference speed (20 fps), allowing for more downstream interactive applications. In addition to strong performance on the standard benchmark, DICE also achieves superior generalization performance on in-the-wild data.

REFERENCES

Oswald Aldrian and William AP Smith. Inverse rendering of faces with a 3d morphable model. *IEEE transactions on pattern analysis and machine intelligence*, 35(5):1080–1093, 2012.

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European conference on computer vision*, pp. 561–578. Springer, 2016.

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

Zhongang Cai, Daxuan Ren, Ailing Zeng, Zhengyu Lin, Tao Yu, Wenjia Wang, Xiangyu Fan, Yang Gao, Yifan Yu, Liang Pan, et al. Humman: Multi-modal 4d human dataset for versatile sensing and modeling. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pp. 557–577. Springer, 2022.

Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12417–12426, 2021.

Zenghao Chai, Tianke Zhang, Tianyu He, Xu Tan, Tadas Baltrusaitis, HsiangTao Wu, Runnan Li, Sheng Zhao, Chun Yuan, and Jiang Bian. Hiface: High-fidelity 3d face reconstruction by learning static and dynamic details. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9087–9098, 2023.

Aggelina Chatziagapi and Dimitris Samaras. Avface: Towards detailed audio-visual 4d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16878–16889, 2023.

Jiayi Chen, Mi Yan, Jiazhao Zhang, Yinzhen Xu, Xiaolong Li, Yijia Weng, Li Yi, Shuran Song, and He Wang. Tracking and reconstructing hand object interactions from point cloud sequences in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 304–312, 2023.

Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE Transactions on Image Processing*, 30:4008–4021, 2021.

Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *European Conference on Computer Vision*, pp. 342–359. Springer, 2022.

Henry M Clever, Patrick L Grady, Greg Turk, and Charles C Kemp. Bodypressure-inferring body pose and contact pressure from a depth image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):137–153, 2022.

MMHuman3D Contributors. Openmmlab 3d human parametric model toolbox and benchmark.

Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5031–5041, 2020.

Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20311–20322, 2022.

Zhiyang Dou, Xuelin Chen, Qingnan Fan, Taku Komura, and Wenping Wang. C· ase: Learning conditional adversarial skill embeddings for physics-based characters. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023a.

Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, and Wenping Wang. Tore: Token reduction for efficient human mesh recovery with transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15143–15155, 2023b.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Qi Fang, Qing Shuai, Junting Dong, Hujun Bao, and Xiaowei Zhou. Reconstructing 3d human pose by watching humans in the mirror. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12814–12823, 2021.

Yao Feng, Fan Wu, Xiaohu Shao, Yanfeng Wang, and Xi Zhou. Joint 3d face reconstruction and dense alignment with position map regression network. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 534–551, 2018.

Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pp. 792–804, 2021a.

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

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Posegpt: Chatting about 3d human pose. *arXiv preprint arXiv:2311.18836*, 2023.

Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1471–1481, 2021.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5152–5161, 2022.

Yong Guo, Zhiyang Dou, Nan Zhang, Xiyue Liu, Boni Su, Yuguo Li, and Yinping Zhang. Student close contact behavior and covid-19 transmission in china's classrooms. *PNAS nexus*, 2(5):pgad142, 2023.

Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3196–3206, 2020.

Xuchen Han, Theodore F. Gast, Qi Guo, Stephanie Wang, Chenfanfu Jiang, and Joseph Teran. A hybrid material point method for frictional contact with diverse materials. 2(2), 2019. doi: 10.1145/3340258. URL `https://doi.org/10.1145/3340258`.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2282–2292, 2019.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14708–14718, 2021.

Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11807–11816, 2019.

Shan He, Haonan He, Shuo Yang, Xiaoyan Wu, Pengcheng Xia, Bing Yin, Cong Liu, Lirong Dai, and Chang Xu. Speech4mesh: Speech-assisted monocular 3d facial reconstruction for speech-driven 3d facial animation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14192–14202, 2023.

Liwen Hu, Shunsuke Saito, Lingyu Wei, Koki Nagano, Jaewoo Seo, Jens Fursund, Iman Sadeghi, Carrie Sun, Yen-Chun Chen, and Hao Li. Avatar digitization from a single image for real-time rendering. *ACM Transactions on Graphics (ToG)*, 36(6):1–14, 2017.

Yuanming Hu, Yu Fang, Ziheng Ge, Ziyin Qu, Yixin Zhu, Andre Pradhana, and Chenfanfu Jiang. A moving least squares material point method with displacement discontinuity and two-way rigid body coupling. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018.

Yuanming Hu, Tzu-Mao Li, Luke Anderson, Jonathan Ragan-Kelley, and Frédo Durand. Taichi: a language for high-performance computation on spatially sparse data structures. *ACM Transactions on Graphics (TOG)*, 38(6):201, 2019.

Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6417–6426, 2022a.

Chun-Hao P Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13274–13285, 2022b.

Kemeng Huang, Floyd M. Chitalu, Huancheng Lin, and Taku Komura. Gipc: Fast and stable gauss-newton optimization of ipc barrier energy. 43(2), 2024. ISSN 0730-0301. doi: 10.1145/3643028. URL `https://doi.org/10.1145/3643028`.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7122–7131, 2018.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Muhammed Kocabas, Chun-Hao P Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J Black. Spec: Seeing people in the wild with an estimated camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11035–11045, 2021.

Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2252–2261, 2019a.

Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4501–4510, 2019b.

Raja Kumar, Jiahao Luo, Alex Pang, and James Davis. Disjoint pose and shape for 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3115–3125, 2023.

Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J Black, and Peter V Gehler. Unite the people: Closing the loop between 3d and 2d human representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6050–6059, 2017.

Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, and Adam Kortylewski. Robust model-based face reconstruction through weakly-supervised outlier segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 372–381, 2023a.

Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3383–3393, 2021.

Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12933–12942, 2023b.

Jiefeng Li, Siyuan Bian, Chao Xu, Zhicun Chen, Lixin Yang, and Cewu Lu. Hybrik-x: Hybrid analytical-neural inverse kinematics for whole-body mesh recovery. *arXiv preprint arXiv:2304.05690*, 2023c.

Kailin Li, Lixin Yang, Haoyu Zhen, Zenan Lin, Xinyu Zhan, Licheng Zhong, Jian Xu, Kejian Wu, and Cewu Lu. Chord: Category-level hand-held object reconstruction via shape deformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9444–9454, 2023d.

Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2761–2770, 2022a.

Minchen Li, Zachary Ferguson, Teseo Schneider, Timothy Langlois, Denis Zorin, Daniele Panozzo, Chenfanfu Jiang, and Danny M. Kaufman. Incremental potential contact: intersection-and inversion-free, large-deformation dynamics. 39(4), 2020. ISSN 0730-0301. doi: 10.1145/3386569.3392425. URL https://doi.org/10.1145/3386569.3392425.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Trans. Graph.*, 36(6):194–1, 2017.

Zhi Li, Soshi Shimada, Bernt Schiele, Christian Theobalt, and Vladislav Golyanik. Mocapdeform: Monocular 3d human motion capture in deformable scenes. In *2022 International Conference on 3D Vision (3DV)*, pp. 1–11. IEEE, 2022b.

Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, pp. 590–606. Springer, 2022c.

Huancheng Lin, Floyd M. Chitalu, and Taku Komura. Isotropic arap energy using cauchy-green invariants. 41(6), 2022. ISSN 0730-0301. doi: 10.1145/3550454.3555507. URL https://doi.org/10.1145/3550454.3555507.

Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, and Yu Li. One-stage 3d whole-body mesh recovery with component aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21159–21168, 2023.

Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1954–1963, 2021a.

Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12939–12948, 2021b.

Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14687–14697, 2021.

Xiyue Liu, Zhiyang Dou, Lei Wang, Boni Su, Tianyi Jin, Yong Guo, Jianjian Wei, and Nan Zhang. Close contact behavior-based covid-19 transmission and interventions in a subway system. *Journal of Hazardous Materials*, 436:129233, 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pp. 851–866. 2023.

I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, et al. Mediapipe: A framework for perceiving and processing reality. In *Third workshop on computer vision for AR/VR at IEEE computer vision and pattern recognition (CVPR)*, volume 2019, 2019.

Zhengyi Luo, Shun Iwase, Ye Yuan, and Kris Kitani. Embodied scene-aware human pose estimation. *Advances in Neural Information Processing Systems*, 35:6815–6828, 2022.

Gyeongsik Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17028–17037, 2023.

Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision*, pp. 752–768. Springer, 2020.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pp. 548–564. Springer, 2020.

Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mallorie Mize, Philippe De Bree, et al. A dataset of relighted 3d interacting hands. *Advances in Neural Information Processing Systems*, 36, 2024.

Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (ToG)*, 38(4): 1–13, 2019.

Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9990–9999, 2021.

Yeonguk Oh, JoonKyu Park, Jaeha Kim, Gyeongsik Moon, and Kyoung Mu Lee. Recovering 3d hand mesh sequence from a single blurry image: A new dataset and temporal unfolding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 554–563, 2023.

Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 1862–1869. IEEE, 2012.

Christopher Otto, Prashanth Chandran, Gaspard Zoss, Markus Gross, Paulo Gotardo, and Derek Bradley. A perceptual shape loss for monocular 3d face reconstruction. In *Computer Graphics Forum*, volume 42, pp. e14945. Wiley Online Library, 2023.

Dongwei Pan, Long Zhuo, Jingtan Piao, Huiwen Luo, Wei Cheng, Yuxin Wang, Siming Fan, Shengqi Liu, Lei Yang, Bo Dai, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, and Kwan-Yee Lin. Renderme-360: Large digital asset library and benchmark towards high-fidelity head avatars. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023a.

Liang Pan, Jingbo Wang, Buzhen Huang, Junyu Zhang, Haofan Wang, Xu Tang, and Yangang Wang. Synthesizing physically plausible human motions in 3d scenes. *arXiv preprint arXiv:2308.09036*, 2023b.

JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1496–1505, 2022.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10975–10985, 2019.

Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (ToG)*, 40(4):1–20, 2021.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions On Graphics (TOG)*, 41(4):1–17, 2022.

Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2883–2896, 2017.

Chandradeep Pokhariya, Ishaan N Shah, Angela Xing, Zekun Li, Kefan Chen, Avinash Sharma, and Srinath Sridhar. Manus: Markerless hand-object grasp capture using articulated 3d gaussians. *arXiv preprint arXiv:2312.02137*, 2023.

Albert Pumarola, Antonio Agudo, Aleix M Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. Ganimation: Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 818–833, 2018.

Dafei Qin, Jun Saito, Noam Aigerman, Thibault Groueix, and Taku Komura. Neural face rigging for animating and retargeting facial meshes in the wild. In *ACM SIGGRAPH 2023 Conference Proceedings*, pp. 1–11, 2023.

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

Davis Rempe, Leonidas J Guibas, Aaron Hertzmann, Bryan Russell, Ruben Villegas, and Jimei Yang. Contact and human dynamics from monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pp. 71–87. Springer, 2020.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11488–11499, October 2021.

Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *2021 International Conference on 3D Vision (3DV)*, pp. 432–441. IEEE, 2021.

Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7763–7772, 2019.

Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, and Jungdam Won. Phasemp: Robust 3d pose estimation via phase-conditioned human motion prior. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14725–14737, 2023.

Soshi Shimada, Vladislav Golyanik, Zhi Li, Patrick Pérez, Weipeng Xu, and Christian Theobalt. Hulc: 3d human motion capture with pose manifold sampling and dense contact guidance. In *European Conference on Computer Vision*, pp. 516–533. Springer, 2022.

Soshi Shimada, Vladislav Golyanik, Patrick Pérez, and Christian Theobalt. Decaf: Monocular deformation capture for face and hand interactions. *ACM Transactions on Graphics (TOG)*, 42(6): 1–16, 2023.

Jan L Spille, Martin Grunwald, Sören Martin, and Stefanie M Mueller. Stop touching your face! a systematic review of triggers, characteristics, regulatory functions and neuro-physiology of facial self-touch. *Neuroscience & Biobehavioral Reviews*, 128:102–116, Sep 2021. doi: 10.1016/j.neubiorev.2021.05.030.

Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019.

Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4511–4520, 2019.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.

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

Edith Tretschk, Navami Kairanda, Mallikarjun BR, Rishabh Dabral, Adam Kortylewski, Bernhard Egger, Marc Habermann, Pascal Fua, Christian Theobalt, and Vladislav Golyanik. State of the art in dense monocular non-rigid 3d reconstruction. In *Computer Graphics Forum*, volume 42, pp. 485–520. Wiley Online Library, 2023.

Aggeliki Tsoli and Antonis A Argyros. Joint 3d tracking of a deformable object in interaction with a hand. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 484–500, 2018.

Weilin Wan, Zhiyang Dou, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Tlcontrol: Trajectory and language control for human motion synthesis. *arXiv preprint arXiv:2311.17135*, 2023a.

Weilin Wan, Yiming Huang, Shutong Wu, Taku Komura, Wenping Wang, Dinesh Jayaraman, and Lingjie Liu. Diffusionphase: Motion diffusion in frequency domain. *arXiv preprint arXiv:2312.04036*, 2023b.

Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020.

Jingbo Wang, Yu Rong, Jingyuan Liu, Sijie Yan, Dahua Lin, and Bo Dai. Towards diverse and natural scene-aware 3d human motion synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20460–20469, 2022a.

Lizhen Wang, Zhiyuan Chen, Tao Yu, Chenguang Ma, Liang Li, and Yebin Liu. Faceverse: a fine-grained and detail-controllable 3d face morphable model from a hybrid dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20333–20342, 2022b.

Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, and Taku Komura. Zolly: Zoom focal length correctly for perspective-distorted human mesh reconstruction. *arXiv preprint arXiv:2303.13796*, 2023a.

Yanjun Wang, Qingping Sun, Wenjia Wang, Jun Ling, Zhongang Cai, Rong Xie, and Li Song. Learning dense uv completion for human mesh recovery. *arXiv preprint arXiv:2307.11074*, 2023b.

Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14644–14654, 2023.

Shih-En Wei, Jason Saragih, Tomas Simon, Adam W Harley, Stephen Lombardi, Michal Perdoch, Alexander Hypes, Dawei Wang, Hernan Badino, and Yaser Sheikh. Vr facial animation via multiview image translation. *ACM Transactions on Graphics (TOG)*, 38(4):1–16, 2019.

Zhenzhen Weng and Serena Yeung. Holistic 3d human and scene mesh estimation from single view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 334–343, 2021.

Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Matthew Johnson, Jingjing Shen, Nikola Milosavljević, Daniel Wilde, Stephan Garbin, Toby Sharp, Ivan Stojiljković, et al. 3d face reconstruction with dense landmarks. In *European Conference on Computer Vision*, pp. 160–177. Springer, 2022.

Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Chore: Contact, human and object reconstruction from a single rgb image. In *European Conference on Computer Vision*, pp. 125–145. Springer, 2022.

Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14928–14940, 2023.

Sirui Xu, Ziyin Wang, Yu-Xiong Wang, and Liang-Yan Gui. Interdreamer: Less supervision for more generalizable text-driven 3d human-object interaction synthesis. In *NeurIPS*, 2024.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11097–11106, 2021.

Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2750–2760, 2022a.

Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20953–20962, 2022b.

Xueze Yang, Zhiyang Dou, Yuqing Ding, Boni Su, Hua Qian, and Nan Zhang. Analysis of sars-cov-2 transmission in airports based on real human close contact behaviors. *Journal of Building Engineering*, 82:108299, 2024.

Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3895–3905, 2022.

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023a.

Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *CVPR*, 2023b.

Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P Breckon, and Jue Wang. Acr: Attention collaboration-based regressor for arbitrary two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12955–12964, 2023.

Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11354–11363, 2021a.

Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11446–11456, 2021b.

Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 34–51. Springer, 2020a.

Nan Zhang, Li Liu, Zhiyang Dou, Xiyue Liu, Xueze Yang, Doudou Miao, Yong Guo, Silan Gu, Yuguo Li, Hua Qian, et al. Close contact behaviors of university and school students in 10 indoor environments. *Journal of Hazardous Materials*, 458:132069, 2023a.

Nan Zhang, Xiyue Liu, Shuyi Gao, Boni Su, and Zhiyang Dou. Popularization of high-speed railway reduces the infection risk via close contact route during journey. *Sustainable Cities and Society*, 99:104979, 2023b.

Nan Zhang, Xueze Yang, Boni Su, and Zhiyang Dou. Analysis of sars-cov-2 transmission in a university classroom based on real human close contact behaviors. *Science of The Total Environment*, 917:170346, 2024.

Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11343–11353, 2021c.

Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, et al. Accurate 3d face reconstruction with facial component tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9033–9042, 2023c.

Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6194–6204, 2020b.

Qingcheng Zhao, Pengyu Long, Qixuan Zhang, Dafei Qin, Han Liang, Longwen Zhang, Yingliang Zhang, Jingyi Yu, and Lan Xu. Media2face: Co-speech facial animation generation with multi-modality guidance. *arXiv preprint arXiv:2401.15687*, 2024.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C Bühler, Xu Chen, Michael J Black, and Otmar Hilliges. Im avatar: Implicit morphable head avatars from videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13545–13555, 2022.

Wenyang Zhou, Zhiyang Dou, Zeyu Cao, Zhouyingcheng Liao, Jingbo Wang, Wenjia Wang, Yuan Liu, Taku Komura, Wenping Wang, and Lingjie Liu. Emdm: Efficient motion diffusion model for fast, high-quality motion generation. *arXiv preprint arXiv:2312.02256*, 2023.

Wojciech Zielonka, Timo Bolkart, and Justus Thies. Towards metrical reconstruction of human faces. In *European Conference on Computer Vision*, pp. 250–269. Springer, 2022.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 813–822, 2019.

Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, and Yangang Wang. Reconstructing interacting hands with interaction prior from monocular images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9054–9064, 2023.

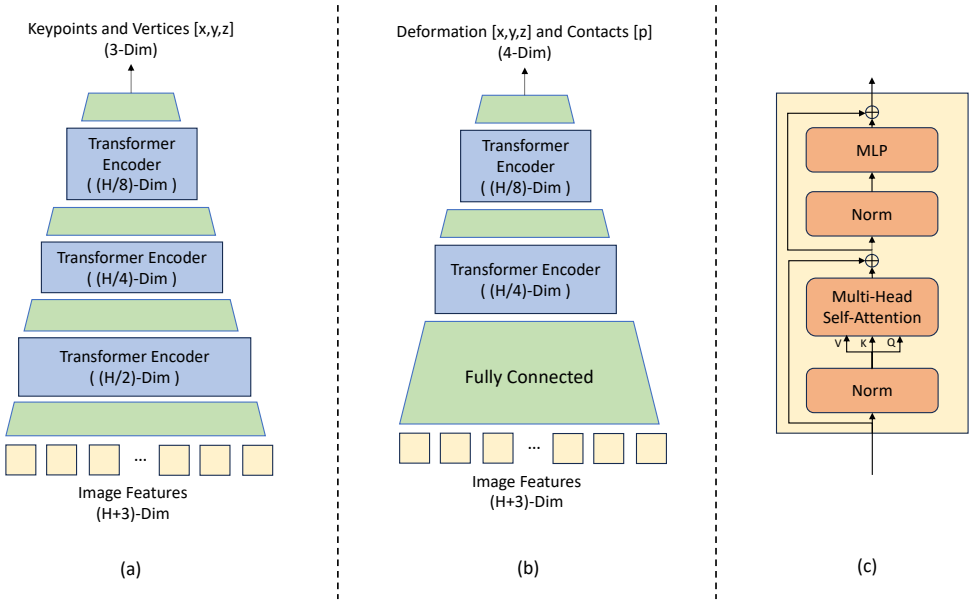


Figure 5: Structural details of the MeshNet and InteractionNet. (a) MeshNet; (b) InteractionNet; (c) Internal structure of a Transformer Encoder block.

A Implementation Details

A.1 CNN Backbone

The CNN backbone used in our framework is an HRNet-W64 (Sun et al., 2019), initialized with ImageNet-pretrained weights. The weights of the backbone would be updated during training. We extract a $(49 \times H)$-dim feature map from this network and upsamples it to a $(N \times H)$-dim feature map, where $N = N_{h_k} + N_{f_k} + N_{h_v} + N_{h_v}$, the total number of head and hand keypoints N_{h_k}, N_{f_k} and vertices N_{h_v}, N_{f_v}. Then, we concatenate the keypoints and the vertices corresponding to the head and hand mean pose as keypoints and vertex queries, resulting in a $((N + 3) \times H)$-dim feature map. Random masking of keypoints and vertex queries of rate 30% is applied, following (Lin et al., 2021a).

A.2 MeshNet and InteractionNet

Our MeshNet and InteractionNet have similar progressive downsampling transformer encoder structures, see Fig. 5 for an illustration. The MeshNet has three component transformer encoders with decreasing feature dimensions. The InteractionNet starts with a fully connected layer that downsamples the feature dimension, followed by two transformer encoders. Each transformer encoder has a Multi-Head Attention module consisting of 4 layers and 4 attention heads. In addition to head and hand mesh features, MeshNet also regresses head and hand keypoints, which are only for supervision and not used by any downstream components.

A.3 IKNet

Our IKNets take in rough mesh features $\mathbf{V}'_F, \mathbf{V}'_H$ and output the pose and shape parameters (θ, β), as well as the global rotation and translation (R, T). They feature a Multi-Layer Perceptron (MLP) structure, each consisting of five MLP Blocks and a final fully connected layer. Each MLP Block contains a fully connected layer, followed by a batch normalization layer (Ioffe & Szegedy, 2015) and a ReLU activation layer. There are two skip-connections, connecting the output of the first block with the input of the third block, and the output of the third block with the input of the final fully connected layer. See Fig. 6 for an illustration. The hand and head IKNets have the same structure, differing only in their input and output dimensions. The hidden dimensions of the two IKNets are 1024.

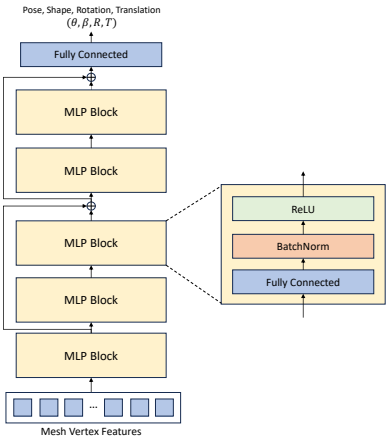


Figure 6: Structural details of the IKNet.

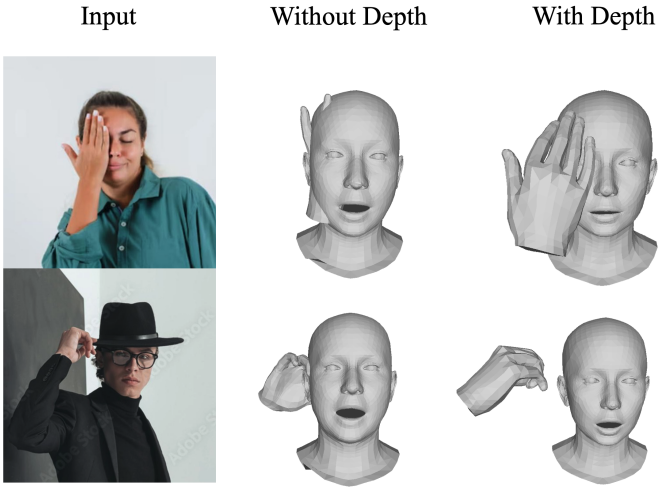


Figure 7: Qualitative demonstration of the effects of the depth loss. The model generalizes poorly in the z-direction when trained without depth supervision.

A.4 Training and Testing Details

To be consistent with the training setting of Decaf[1] (Shimada et al., 2023), in the Decaf dataset, we use all eight camera views and the subjects S2, S4, S5, S7, and S8 in the training data split for training. For testing, we use only the front view (view 108) and the subjects S1, S3, and S6 in the testing data split. The low, mid, and high-resolution head mesh consists of 559, 1675, and 5023 vertices, respectively. The low and high-resolution hand mesh consists of 195 and 778 vertices, respectively. We use the middle-resolution head mesh and the high-resolution hand mesh as the inputs of head and hand IKNets.

B More Qualitative Comparisons

We demonstrate qualitatively the effect of the absence of the depth loss in Fig. 7. When trained without depth loss, the network is only supervised with 2D information on in-the-wild data, without any constraints in the z-direction. As a result, artifacts such as self-penetration frequently occur in this case. The introduction of depth loss eliminates this ambiguity, allowing the correct relative positioning of hand and face.

[1]Confirmed by the authors of Decaf

C ADDITION DETAILS ON LOSSES

Here, we provide the details of the mesh losses and the interaction losses. The details of the adversarial loss and the depth loss are already mentioned in the main paper.

C.1 MESH LOSSES

The mesh loss $\mathcal{L}_{\text{mesh}}$ consists of four components.

$$\mathcal{L}_{\text{mesh}} = \mathcal{L}_{\text{reproj}} + 4\mathcal{L}_{\text{vert}} + 2\mathcal{L}_{\text{key}} + 2\mathcal{L}_{\text{params}} \tag{6}$$

Vertices Loss L_1 loss is used for predicted rough 3D face and hand vertices $\mathbf{V}'_f$, $\mathbf{V}'_h$, FLAME-regressed undeformed 3D face vertices $\mathbf{V}^*_f$ and MANO-regressed 3D hand vertices $\mathbf{V}_h$ against the ground-truth 3D undeformed face vertices $\hat{\mathbf{V}}_f$ and 3D hand vertices $\hat{\mathbf{V}}_h$.

$$\mathcal{L}_{\text{vert}} = \lambda_h(\mu_{\text{nonpara}}\|\mathbf{V}'_h - \hat{\mathbf{V}}_h\|_1 + \|\mathbf{V}_h - \hat{\mathbf{V}}_h\|_1) + \lambda_f(\mu_{\text{nonpara}}\|\mathbf{V}'_f - \hat{\mathbf{V}}_h\|_1 + \|\mathbf{V}^*_f - \hat{\mathbf{V}}_f\|_1) \tag{7}$$

where λ_h, λ_f are empirically set to 3 and 1 respectively. μ_{nonpara} is set to 4 to emphasize the supervision on the more complex non-parametric mesh features.

Keypoints Loss We use L_1 loss for predicted rough 3D face and hand keypoints $\mathbf{K}'_f$, $\mathbf{K}'_h$, 3D face and hand keypoints extracted from rough mesh $\mathbf{K}_{f_{\text{mesh}}}, \mathbf{K}_{h_{\text{mesh}}}$, FLAME-regressed 3D face keypoints $\mathbf{K}_f$ and MANO-regressed 3D hand keypoints $\mathbf{K}_h$ against the ground-truth 3D undeformed face keypoints $\hat{\mathbf{K}}_f$ and 3D hand keypoints $\hat{\mathbf{K}}_f$.

$$\mathcal{L}_{\text{key}} = \mu_{\text{nonpara}}(\|\mathbf{K}'_h - \hat{\mathbf{K}}_h\|_1 + \|\mathbf{K}_{h_{\text{mesh}}} - \hat{\mathbf{K}}_h\|_1 + \|\mathbf{K}'_f - \hat{\mathbf{K}}_f\|_1 + \|\mathbf{K}_{f_{\text{mesh}}} - \hat{\mathbf{K}}_f\|_1) \tag{8}$$

$$+\|\mathbf{K}_f - \hat{\mathbf{K}}_f\|_1 + \|\mathbf{K}_h - \hat{\mathbf{K}}_h\|_1 \tag{9}$$

Where μ_{nonpara} is empirically set to 4, to put more weight on the non-parametric mesh with high degrees of freedom.

Reprojection loss L_1 loss is used for reprojected rough 3D face and hand keypoints $\mathbf{K}'_f$, $\mathbf{K}'_h$, 3D face and hand keypoints extracted from rough mesh $\mathbf{K}_{f_{\text{mesh}}}, \mathbf{K}_{h_{\text{mesh}}}$, FLAME-regressed 3D face keypoints $\hat{\mathbf{K}}_f$ and MANO-regressed 3D hand keypoints $\hat{\mathbf{K}}_h$ against the ground-truth face and hand 2D keypoints $\hat{\mathbf{K}}_{f_{\text{2D}}}, \hat{\mathbf{K}}_{h_{\text{2D}}}$.

$$\mathcal{L}_{\text{reproj}} = \lambda_h(\|\Pi(\mathbf{K}'_h) - \hat{\mathbf{K}}_{h_{\text{2D}}}\|_1 + \|\Pi(\mathbf{K}_{h_{\text{mesh}}}) - \hat{\mathbf{K}}_{h_{\text{2D}}}\|_1 + \|\Pi(\mathbf{K}_h) - \hat{\mathbf{K}}_{h_{\text{2D}}}\|_1) \tag{10}$$

$$+\lambda_f(\|\Pi(\mathbf{K}'_f) - \hat{\mathbf{K}}_{f_{\text{2D}}}\|_1 + \|\Pi(\mathbf{K}_{f_{\text{mesh}}}) - \hat{\mathbf{K}}_{f_{\text{2D}}}\|_1 + \|\Pi(\mathbf{K}_f) - \hat{\mathbf{K}}_{f_{\text{2D}}}\|_1) \tag{11}$$

Where Π is the learned camera projection function. λ_h, λ_f are set to 4 and 1 respectively.

Parameter loss We apply L_1 loss on the regressed hand and face pose, shape, and facial expression parameters against their respective ground truths.

$$\mathcal{L}_{\text{face-params}} = (\|\beta_{\text{f}} - \hat{\beta}_{\text{f}}\|_1 + \|\theta_{\text{f-exp}} - \hat{\theta}_{\text{f-exp}}\|_1 + \|\theta_{\text{f-pose}} - \hat{\theta}_{\text{f-pose}}\|_1)/3 \tag{12}$$

$$\mathcal{L}_{\text{hand-params}} = (\|\beta_{\text{h}} - \hat{\beta}_{\text{h}}\|_1 + \|\theta_{\text{h}} - \hat{\theta}_{\text{h}}\|_1)/2 \tag{13}$$

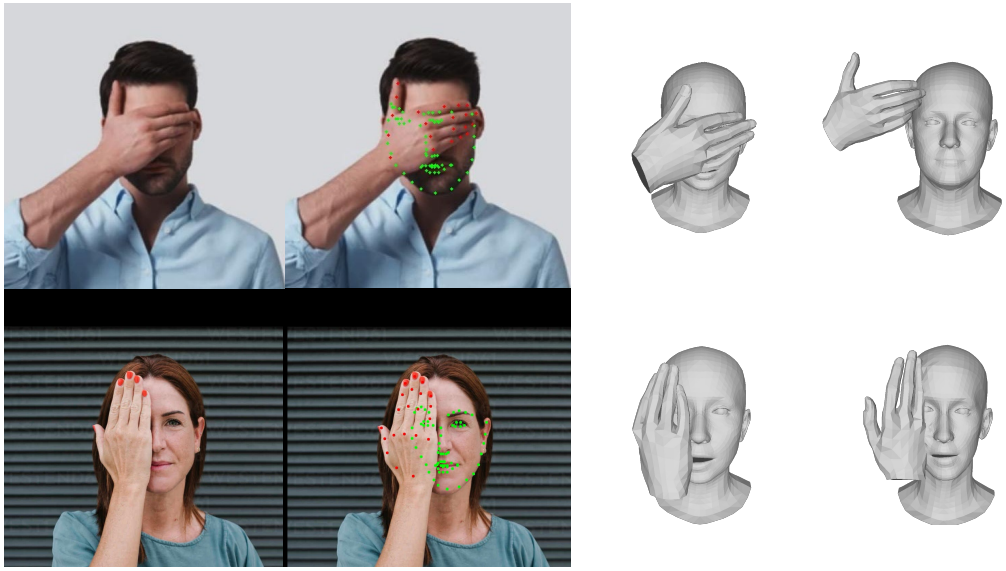
$$\mathcal{L}_{\text{params}} = \mathcal{L}_{\text{face-params}} + \mathcal{L}_{\text{hand-params}} \tag{14}$$

C.2 INTERACTION LOSSES

The interaction loss $\mathcal{L}_{\text{interaction}}$ consists of four components.

$$\mathcal{L}_{\text{interaction}} = 0.2\mathcal{L}_{\text{touch}} + 0.6\mathcal{L}_{\text{contact}} + \mathcal{L}_{\text{collision}} + 6\mathcal{L}_{\text{deform}} \tag{15}$$

Deformation loss Due to the human anatomy, some vertices on the face are more easily deformed than other vertices. Therefore, we impose an adaptive weighting on each vertex, and use square



(a) (b) (c) (d)

Figure 8: Examples of failed keypoint estimation in case of large self-occlusion. (a) input image; (b) inaccurate keypoint estimation by the same keypoint estimators used in Decaf (Lugaresi et al., 2019; Bulat & Tzimiropoulos, 2017); (c) reconstructed hand-face interaction by our method. (d) reconstructed hand-face interaction by Decaf.

loss to penalize large deformation. We also have a regularization term to penalize extremely large deformations.

$$\mathcal{L}_{\text{deform}} = \sum_{i \in \mathcal{I}} (1 + \mu \|\hat{d}_i\|_2) \|\hat{d}_i - d_i\|_2^2 + \lambda \sum_{i \in \mathcal{L}} \|d_i\| \tag{16}$$

Where $\mathcal{I}$ is the set of indices of face vertices, d_i, $\hat{d}_i$ are the predicted and ground truth deformation vector for index i, and $\mathcal{L} = \{i \in \mathcal{I} : \|d_i\|_2 > 3cm\}$ the vertices of large deformations. μ and λ are empirically set to be 5000, 100 respectively.

Touch loss Let $\mathbf{V}_{F_C}$ and $\mathbf{V}_{H_C}$ denote the set of face and hand vertices that are predicted by the model to have contact probability greater than 0.5.

$$\mathcal{L}_{\text{touch}} = \text{CD}(\mathbf{V}_{F_C}, \mathbf{V}_{H_C}) + \text{CD}(\mathbf{V}_{H_C}, \mathbf{V}_{F_C}) \tag{17}$$

Where $\text{CD}(X, Y)$ gives the mean Chamfer Distance (CD) between each point in X to the closest point in Y.

Collision loss Let $\mathbf{V}_{H_{\text{Col}}}$ denote the set of hand vertices that penetrates the face surface, $\mathbf{V}_F$ and $\mathbf{D}_F$ denote the predicted face mesh vertices and deformations.

$$\mathcal{L}_{\text{collision}} = \text{CD}(\mathbf{V}_{H_{\text{Col}}}, \mathbf{V}_F - \mathbf{D}_F) \tag{18}$$

Contact loss Let $\mathbf{C}_H$ and $\mathbf{C}_F$ denote the predicted hand and face contact probabilities and $\hat{\mathbf{C}}_H$, $\hat{\mathbf{C}}_F$ denote the ground-truth contact labels.

$$\mathcal{L}_{\text{contact}} = \text{BCE}(\mathbf{C}_H, \hat{\mathbf{C}}_H) + \text{BCE}(\mathbf{C}_F, \hat{\mathbf{C}}_F) \tag{19}$$

Where BCE denote the binary cross-entropy loss.

D More Discussions

D.1 Performance under Challenging Occlusion.

As seen in Fig. 8, our end-to-end DICE method is robust under challenging self-occlusion cases, such as the hand covering more than half of the face. On the other hand, Decaf (Shimada et al., 2023), which requires an initial keypoint prediction for test-time optimization, performs poorly in this situation.

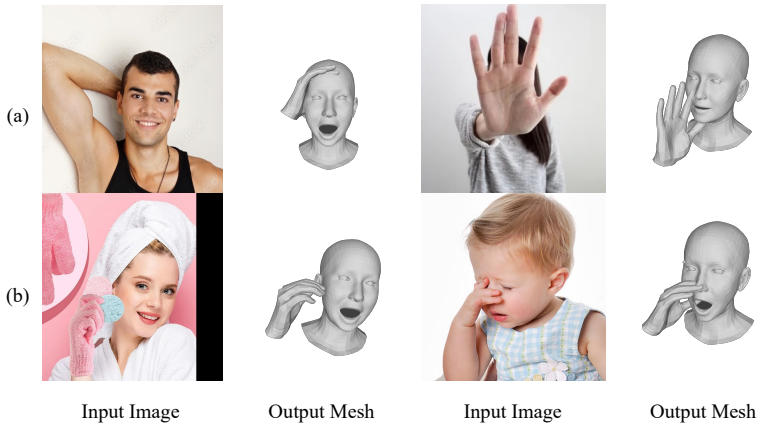


Figure 9: Examples of failure cases in case of complete occlusion of the hand.
(a) Hand or face completely occluded. (b) Out-of-distribution data.

D.2 Failure Cases

In Fig. 9, we demonstrate the failure cases of our method. When the hand is extremely far from the face, or when the hand is completely obscured by the head, our method could fail to reconstruct the hand-face interaction. Also, when given out-of-distribution data, such as when the hand is wearing gloves or the input subject is an infant, the reconstruction accuracy could degrade.

D.3 Societal Impact

D.3.1 Potential Misuse

DICE enables tracking of individuals' appearances, gestures, and interactions with high fidelity, there is a risk that it may be misused for negative applications, such as surveillance, and may cause privacy infringement. Also, since DICE makes use of a readily animatable representation, it could enable realistic deepfakes driven by the pose and shape information collected, which could be used in creating misinformation and conducting identity theft. We are firmly against any form of misuse of the DICE model.

D.3.2 Data Fairness

As hand-face interaction recovery is a human-related task, data fairness is critical. The currently used Decaf Shimada et al. (2023) dataset needs improvement in the inclusion of human actors from underrepresented demographic groups. This may result in a model trained only on Decaf underperforming on input data on such groups, perpetuating inequality and limiting equitable access. Our weak-supervised training scheme introduces diverse in-the-wild data, which could alleviate this issue as the amount of in-the-wild data scales up.