DICE: End-to-end Deformation Capture of Hand-Face Interactions from a Single Image

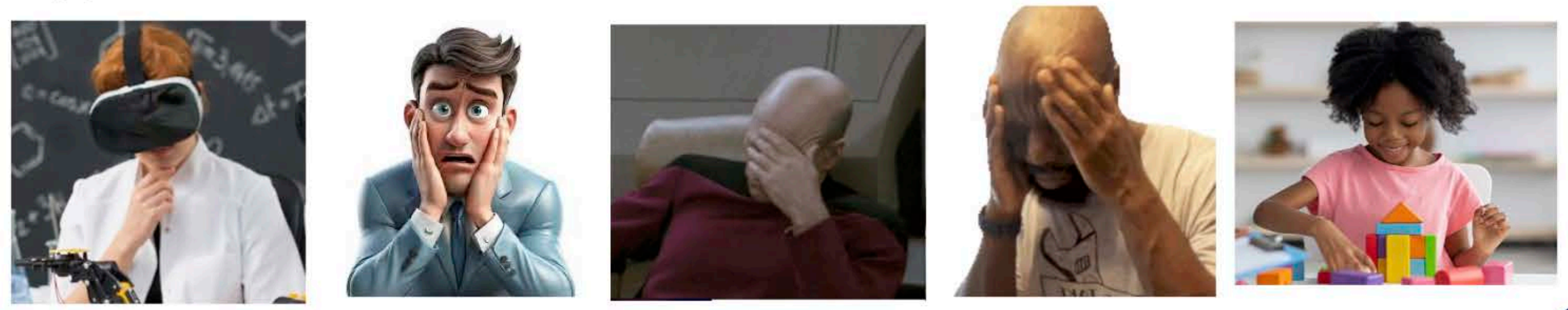
Qingxuan Wu[1], Zhiyang Dou[1,2,7,*] Sirui Xu[3], Soshi Shimada[4], Chen Wang[1], Zhengming Yu[6], Yuan Liu[2], Cheng Lin[2],
Zeyu Cao[5], Taku Komura[2,7], Vladislav Golyanik[4], Christian Theobalt[4], Wenping Wang[6], Lingjie Liu[1,*]
University of Pennsylvania[1], The University of Hong Kong[2], University of Illinois Urbana-Champaign[3], Max Planck Institute for Informatics[4], University of Cambridge[5], Texas A&M University[6], TransGP[7]
(*Corresponding Authors)

ICLR 2025

Video

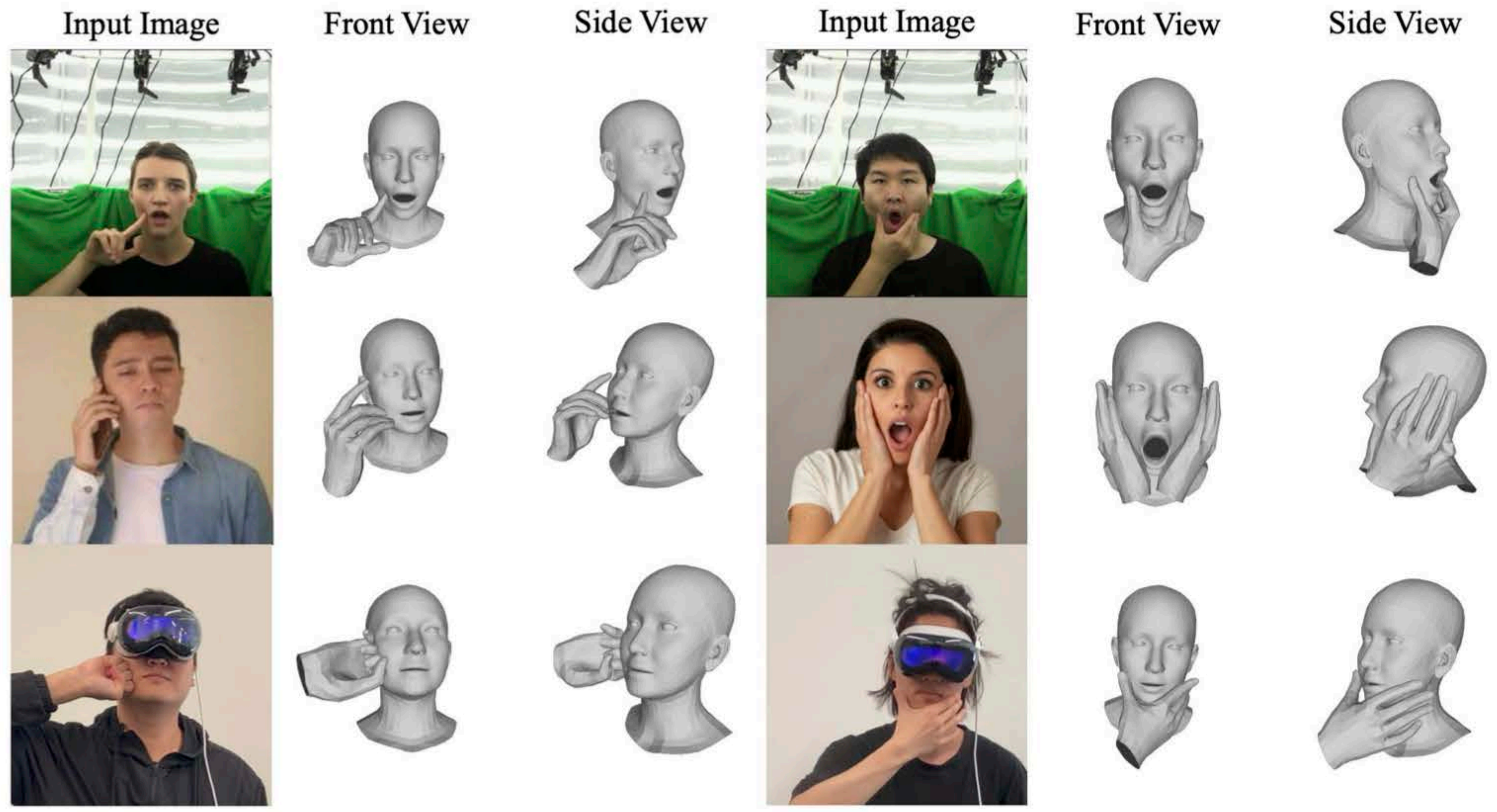
Motivation

- Hand-face interaction is a common behavior observed up to **800 times per day** across all ages and genders.
- Hand-face interaction recovery with deformations has applications in **AR/VR**, **character animation**, and **human behavior analysis**.
- **Accurate + real-time method** required for time-sensitive applications.

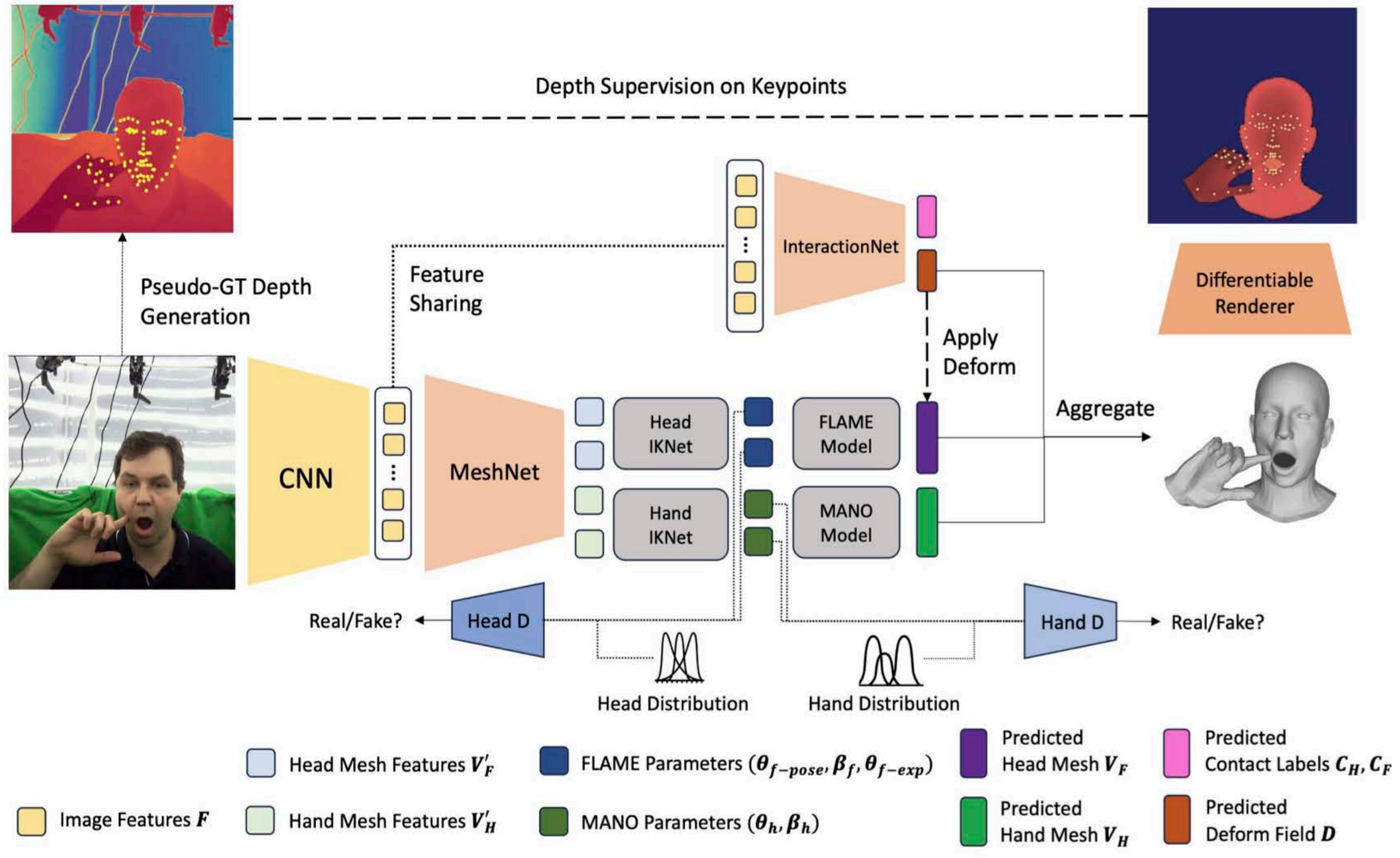


Introduction

- **DICE** is the first **end-to-end** method for Deformation aware hand-face Interaction recovery from a single image.
- DICE achieves **SOTA performance** on standard benchmark and in-the-wild data in terms of accuracy and physical plausibility.
- Additionally, DICE runs at **20 fps** on an Nvidia 4090 GPU, up to **200x faster** than previous methods.



Method



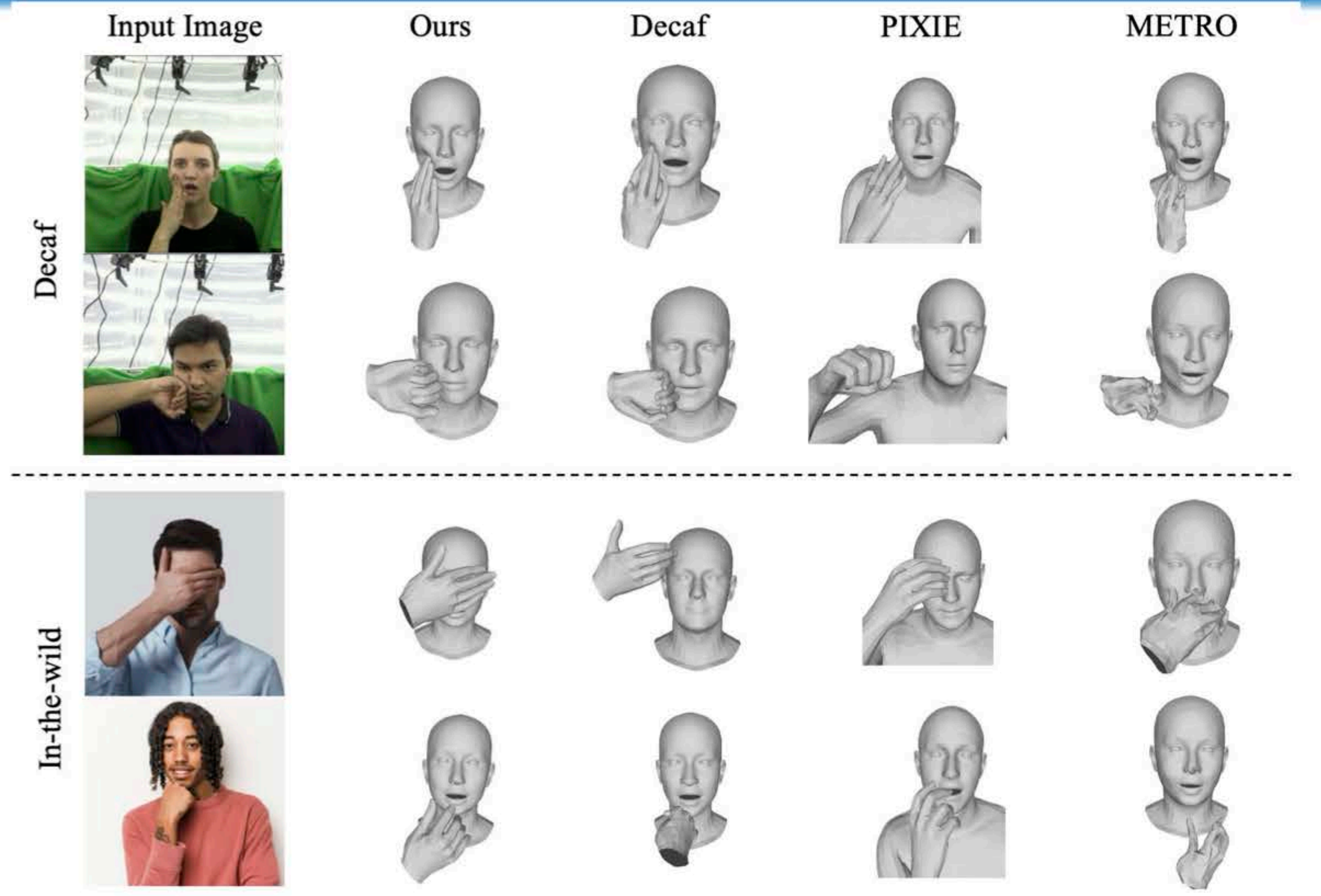
Transformer-based Hand-face Interaction Recovery

- HRNet-W64 CNN extracts feature from single-image input.
- Feature is passed to Transformer-based MeshNet and InteractionNet:
  - **MeshNet** extracts hand and face mesh features;
  - **InteractionNet** predicts per-vertex hand-face contact probabilities and face deform fields.
- The **IKNets** regress FLAME (face) and MANO (hand) parameters as well as rotation/translation parameters.
- Face and hand **mesh** is obtained from FLAME and MANO forward pass.
- The **deformation field** predicted by InteractionNet is applied to the face mesh.

Weakly-Supervised Training Scheme

- **In-the-wild images** are used to enhance generalizing capability.
- **Pseudo-GT depth** and **2D keypoints** are annotated with off-the-shelf models and used for supervision.
- **Adversarial training** with hand-/face- only dataset is employed to constrain the face and hand parameter distribution.

Comparison with Previous Methods



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.

Qualitative Results