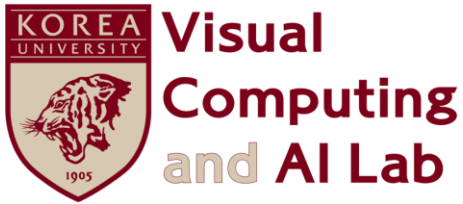
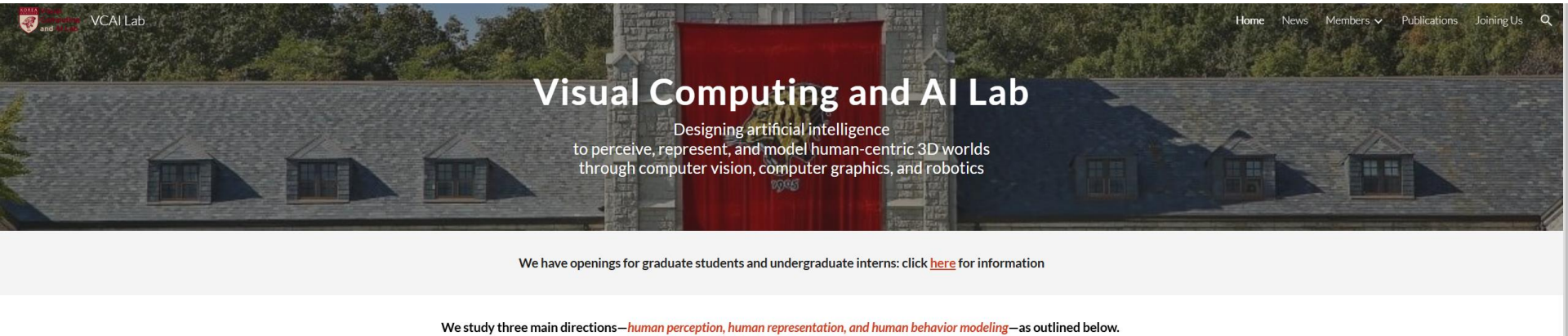
Human-Understanding AI

Gyeongsik Moon

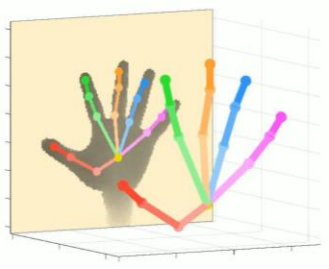


- Gyeongsik Moon (문경식)
- Department of CSE (컴퓨터 학과)
- Computer Vision, Computer Graphics, Robotics
- Fresh faculty (first semester!)
- 1 year at DGIST, 2 years at Meta after finishing Ph.D. at SNU in 2021
- Website: www.vcai.korea.ac.kr

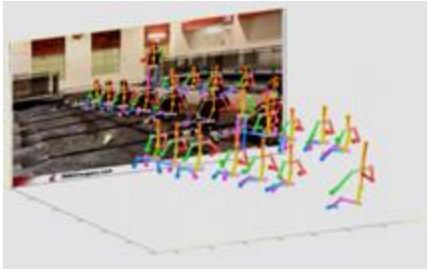


ComputerVisionLab
Seoul National University

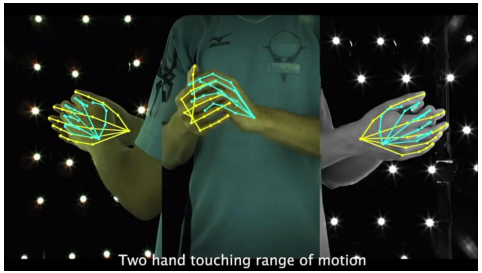
Computer Vision



3D hand pose estimation
(CVPR 2018)



3D multi-person pose estimation
(ICCV 2019)



3D interacting hand pose estimation
(ECCV 2020)

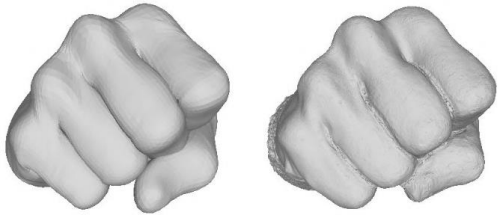
3D body shape estimation
(ECCV 2020)

3D body shape estimation
(CVPR 2021)

Hand-object interaction
(CVPR 2022)

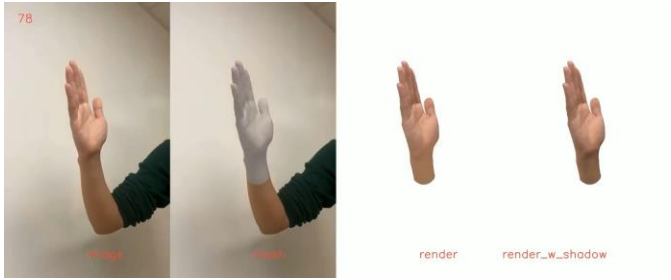
3D whole-body pose estimation
(CVPR 2022)

Computer Graphics

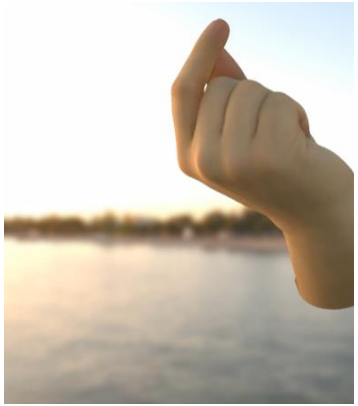


High-fidelity 3D hand geometry model
(ECCV 2020)

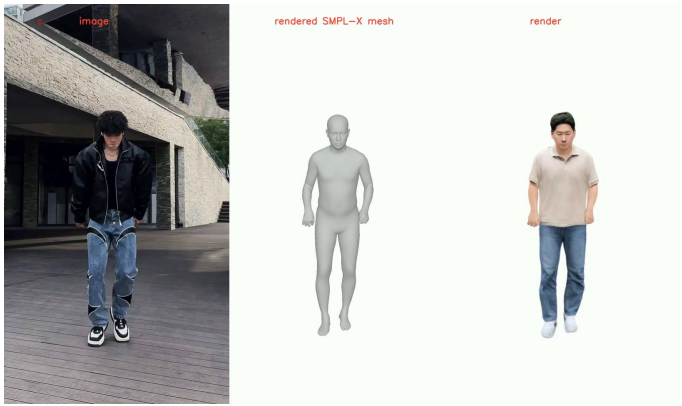
Relighted 3D interacting hands
(NeurIPS 2023)



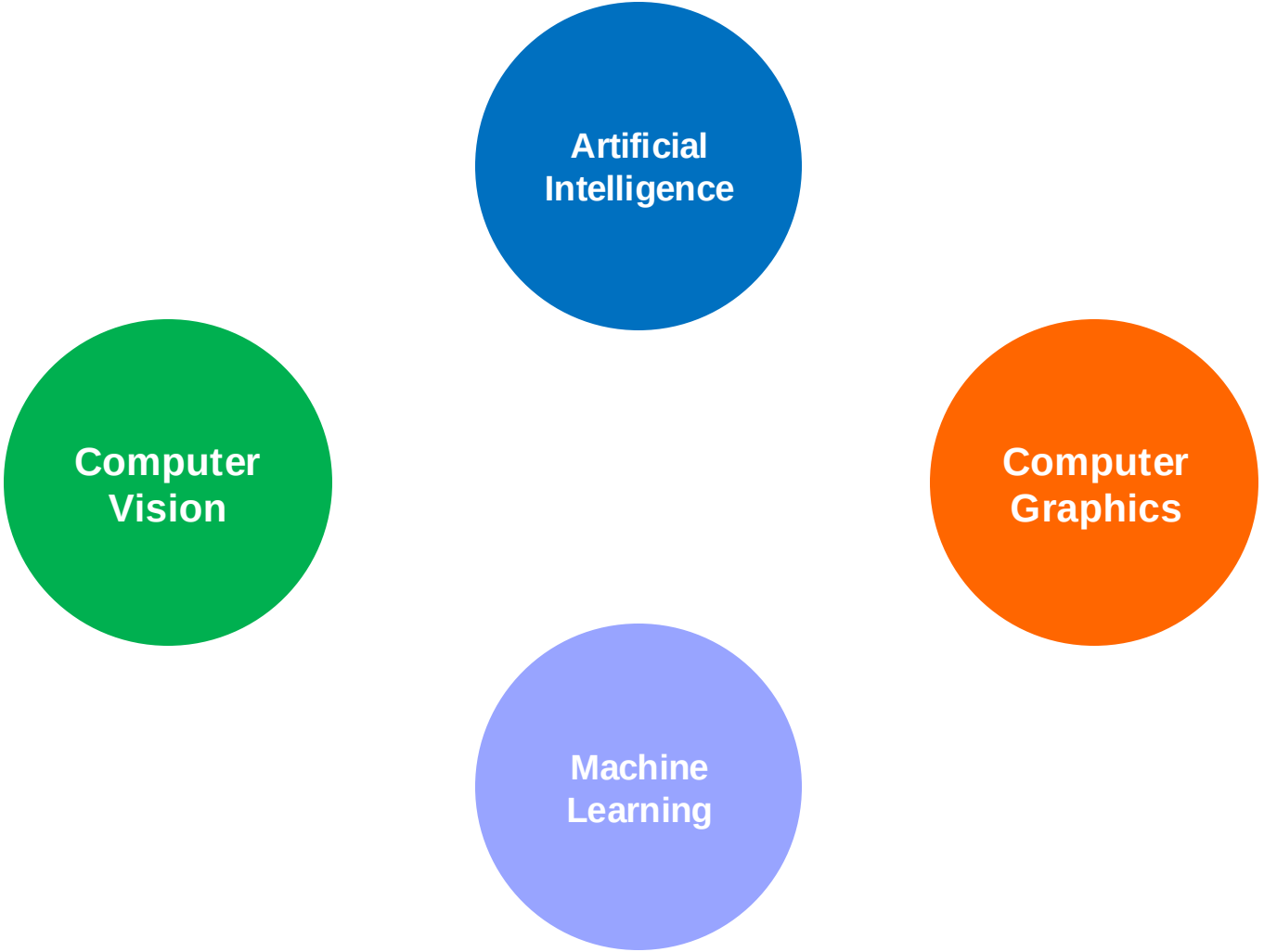
Authentic 3D hand avatar
(CVPR 2024)

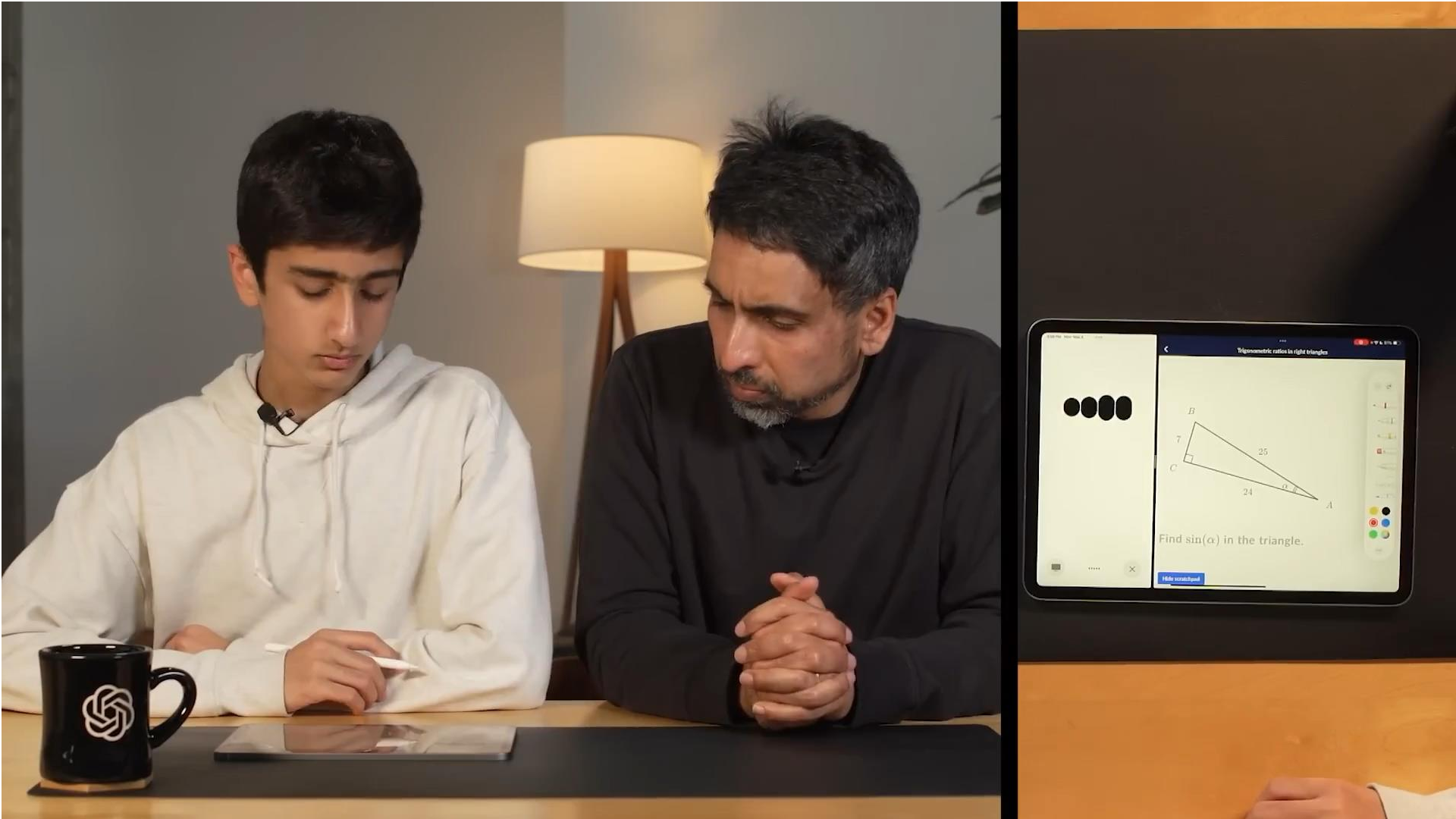


Universal relightable hand model
(CVPR 2024)



Expressive Whole-Body 3D Gaussian Avatar
(ECCV 2024)


Artificial Intelligence
Computer Vision
Computer Graphics
Machine Learning

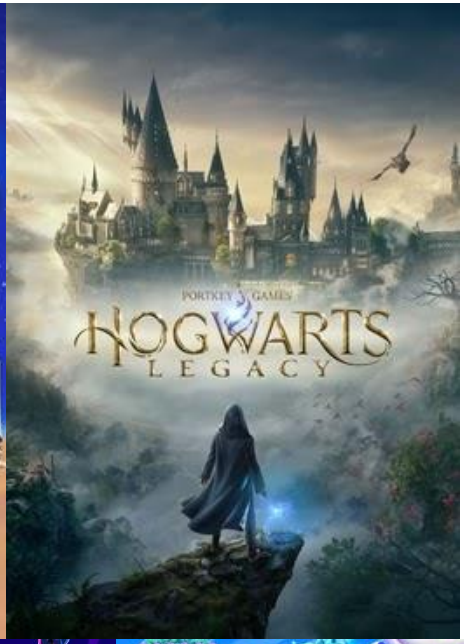

Trigonometric ratios in right triangles
B
7
25
C
24
A
Find sin(α) in the triangle.
Hide scratchpad


Understand humans, and
eventually, interact with humans


BLACK MYTH
WUKONG


龍が
LIKE A DRAGON
INFINITE WEALTH
如


ASTRO
BOT


PORTKEY GAMES
HOGWARTS
LEGACY


LIES OF P


blasphemous II


Stellar
BLADE


SANABI


SUPER
MARIO BROS.
WONDER

Humans are at EVERYWHERE

What makes it challenging?

“A Korean guy is breakdancing” with SORA of OpenAI


LUMA

Human image: Lots of physical attributes

Pose

Facial expression

Cloth dynamics

Viewpoint

Illumination

Human image: Lots of physical attributes

- Human is the most difficult object to represent
- In particular, it is dynamic (e.g., diverse body poses) in contrast to static scenes or objects

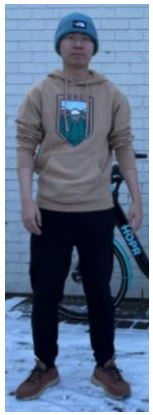
Dynamic humans

>

More challenging

Static scenes

Perceive and Represent Human Images



Perceive
(inverse rendering)

Control
(pose, light, …)

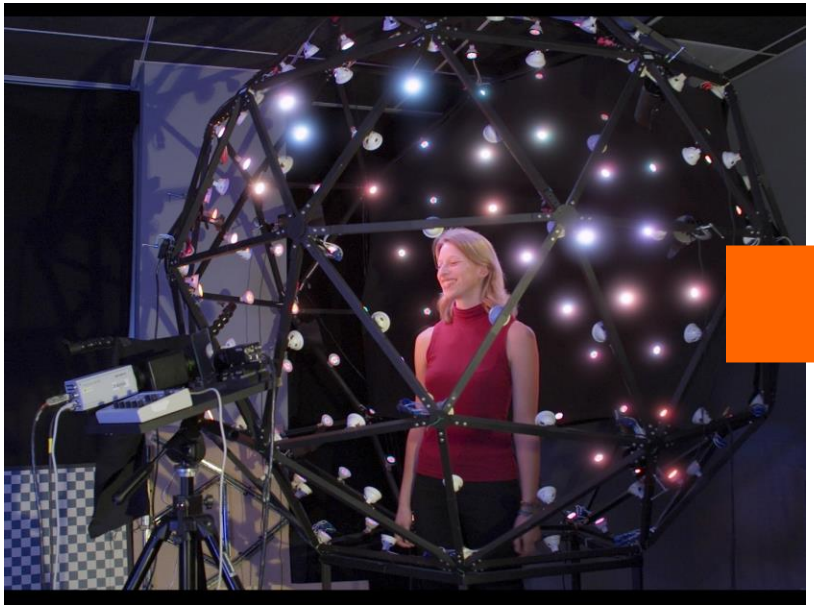
Represent
(rendering)

Human images

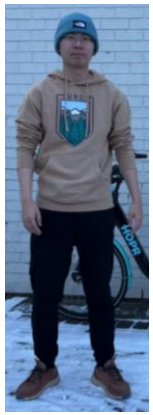
Human features
(e.g., geometry, texture, light, latent features, …)

Understanding Humans in Casual Environments

- Casual environments have much less available observations
- Need to use **priors to complement limited observations**



Perceive and Represent Human Images



Perceive
(inverse rendering)

Control
(pose, light, …)

Represent
(rendering)

Human images

Human features
(e.g., geometry, texture, light, latent features, …)

URHand: Universal Relightable Hands

Zhaoxi Chen[1,2], Gyeongsik Moon[1], Kaiwen Guo[1], Chen Cao[1], Stanislav Pidhorskyi[1]
Tomas Simon[1], Rohan Joshi[1], Yuan Dong[1], Yichen Xu[1], Bernardo Pires[1],
He Wen[1], Lucas Evans[1], Bo Peng[1], Julia Buffalini[1], Autumn Trimble[1],
Kevyn McPhail[1], Melissa Schoeller[1], Shoou-I Yu[1], Javier Romero[1],
Michael Zollhöfer[1], Yaser Sheikh[1], Ziwei Liu[2] , Shunsuke Saito[1]

[1]Codec Avatars Lab, Meta *[2]Nanyang Technological University*

Why Relighting?

Generalize to novel illuminations

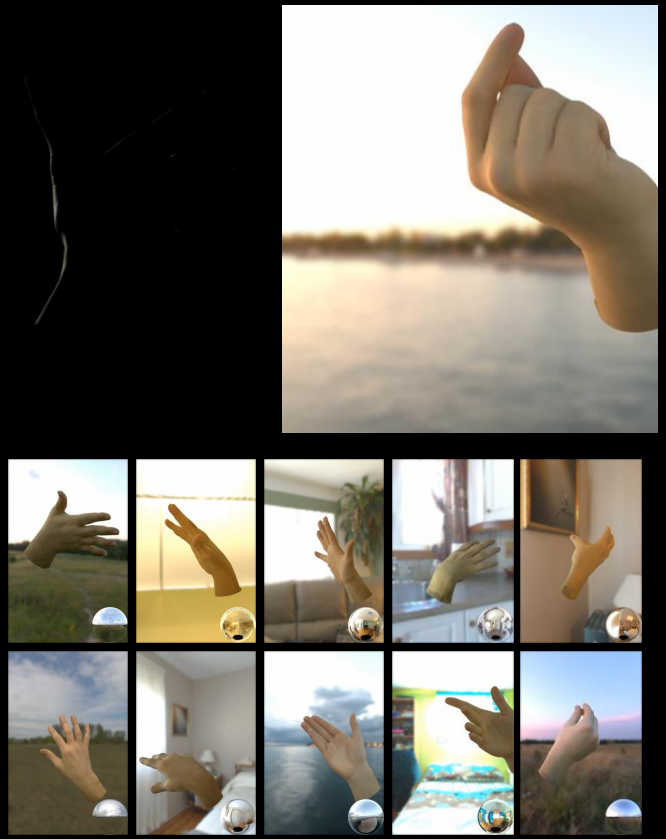
Given large-scale light-stage data…

[1] Chen et al. ” URHand: Universal Relightable Hands.” CVPR. 2024.

[1] Chen et al. ” URHand: Universal Relightable Hands.” CVPR. 2024.

Light-stage Data

URHand: Universal Prior for Relightable Hands

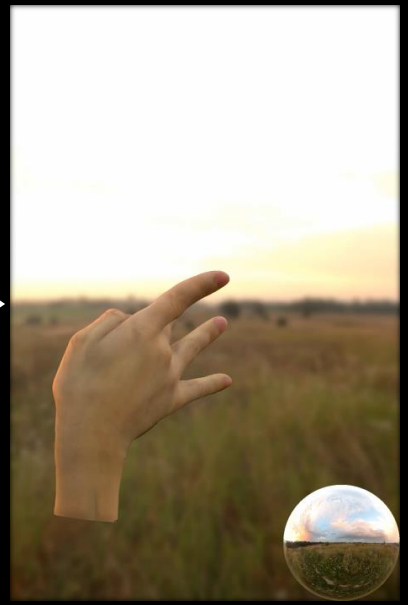


Generalize to Novel Views, Poses, Identities, and Illuminations

[1] Chen et al. ” URHand: Universal Relightable Hands.” CVPR. 2024.

URHand: Universal Prior
for Relightable Hands

[1] Chen et al. ” URHand: Universal Relightable Hands.” CVPR. 2024.

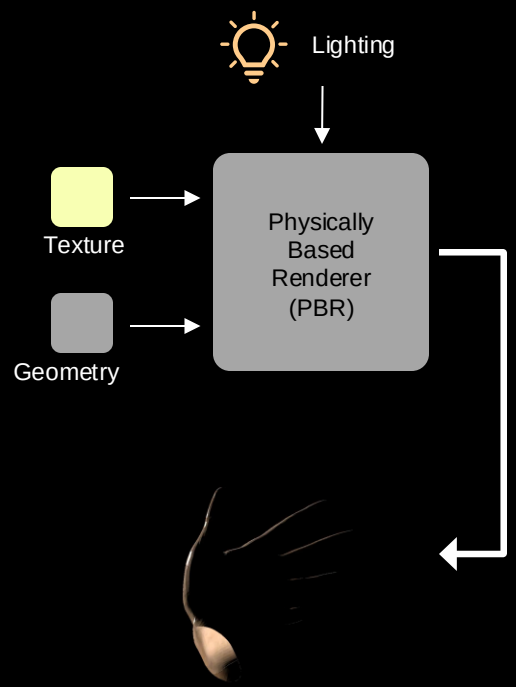


URHand: Universal Prior
for Relightable Hands

Your Hand Relightable

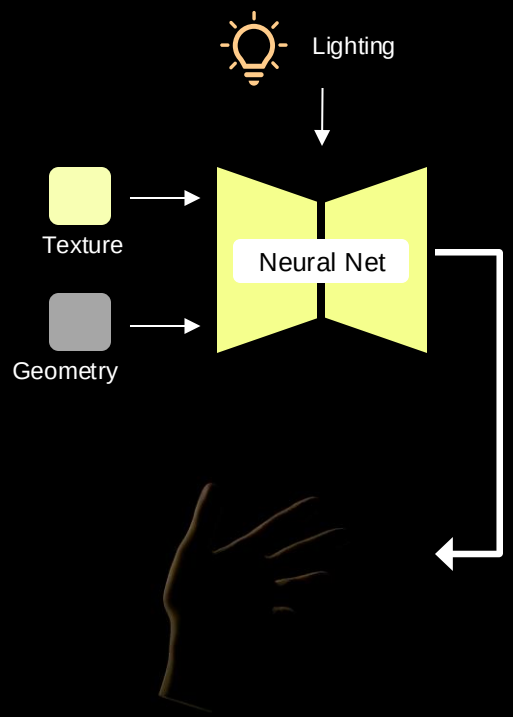
Quick Personalization from a Phone Scan

[1] Chen et al. ” URHand: Universal Relightable Hands.” CVPR. 2024.



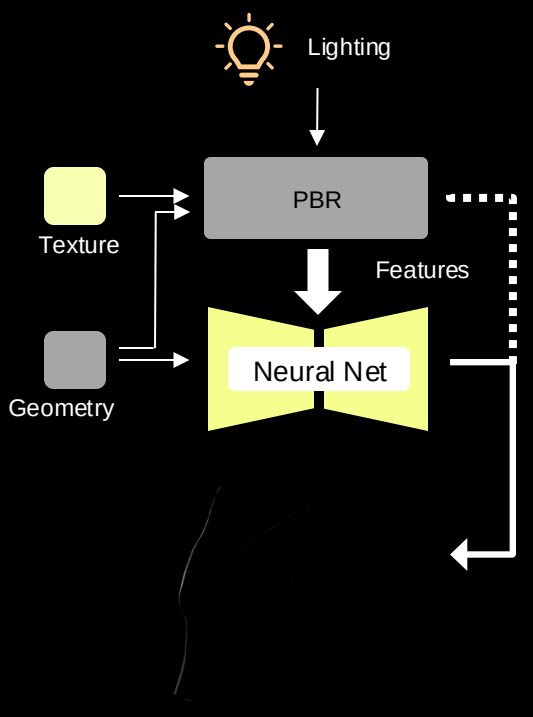
Physically Based Relighting

- ✚ Generalizable in extrapolation
- ▬ Slow synthesis and low photorealism



Neural Relighting

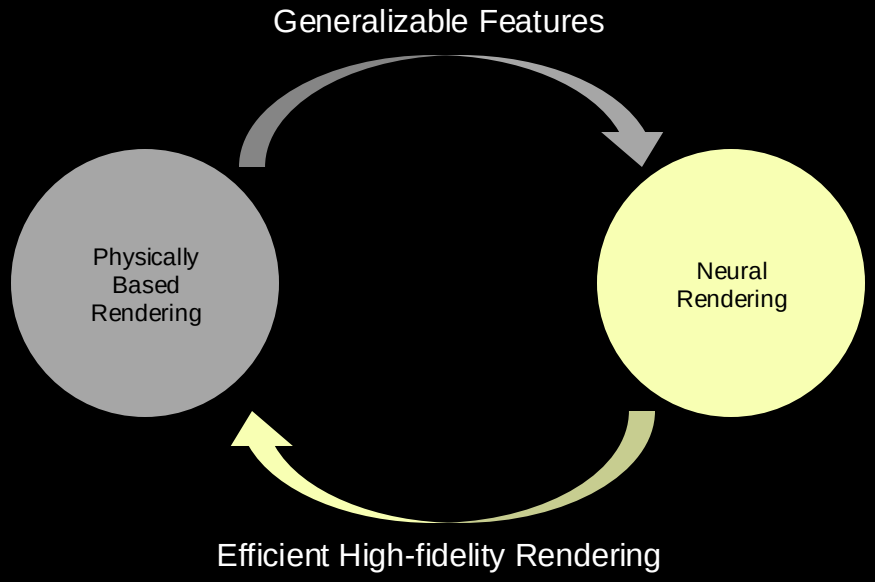
- ✚ Fast synthesis and high photorealism
- ▬ Hard to generalize in extrapolation

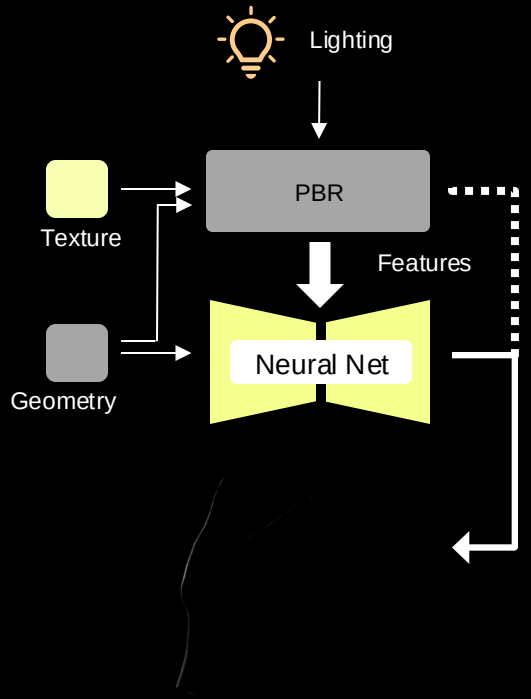


Neural-Physical Relighting

- ✚ Generalizable in extrapolation
- ✚ Fast synthesis and high photorealism

Our Solution:
Neural-Physical Relighting

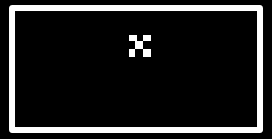




Neural-Physical Relighting

- ✚ Generalizable in extrapolation
- ✚ Fast synthesis and high photorealism

A Gap between Training and Inference



Training

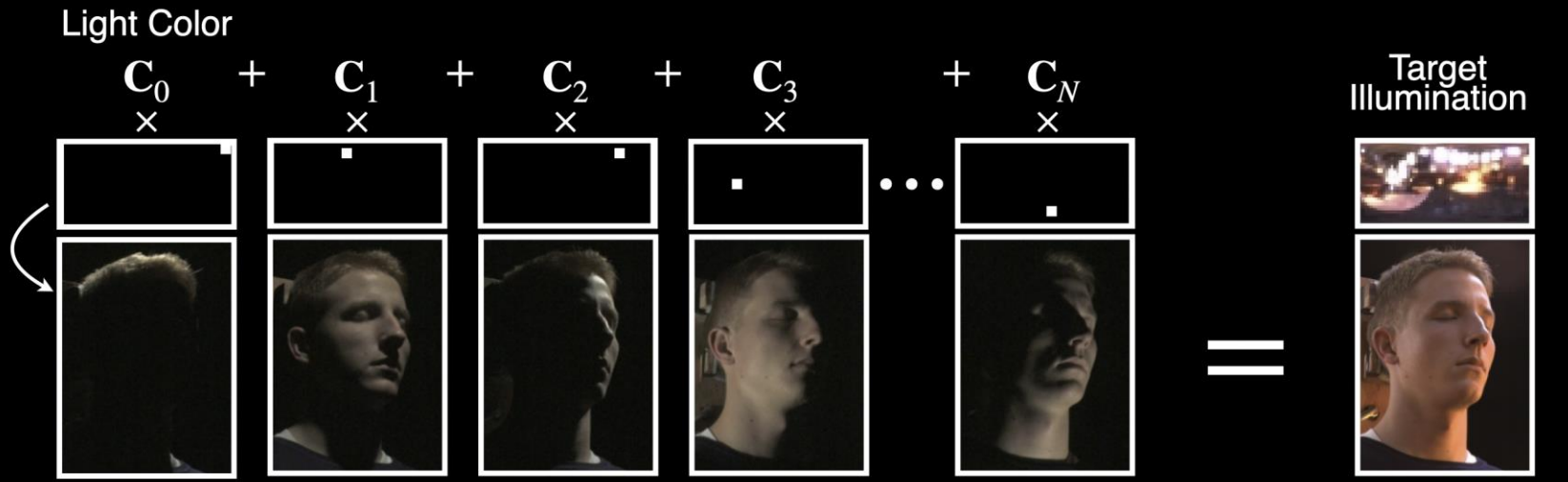
One Light At a Time (OLAT)

Inference

Combination of multiple natural lights

Generalize?

Recap: Linearity of Light Transport



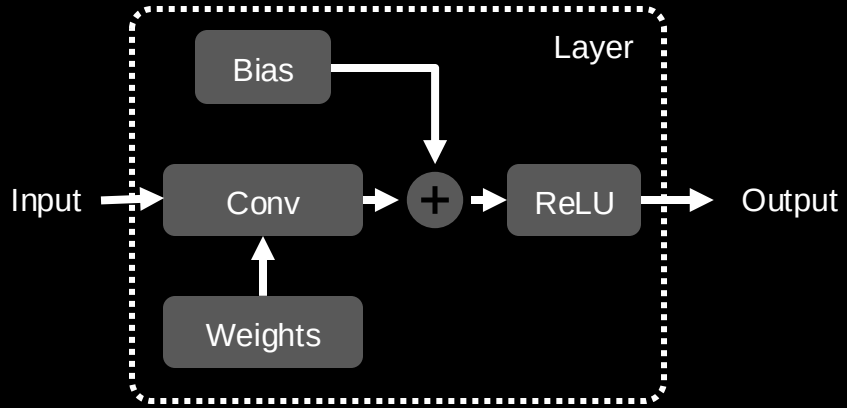
One Light At a Time (OLAT) Images

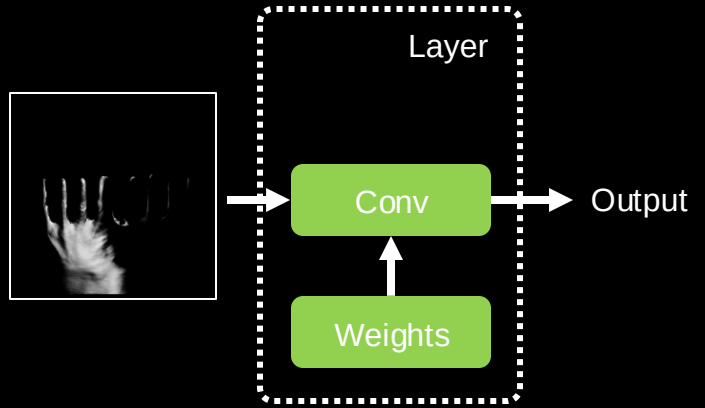
Linear Combinations

[1] Debevec et al. “Acquiring the Reflectance Field of a Human Face”. SIGGRAPH. 2000.

Key Idea: Linear Lighting Model

$$\sum \mathbf{C}_i \boldsymbol{f}_{\mathbf{linear}}(\boldsymbol{b}_i) = \boldsymbol{f}_{\mathbf{linear}}(\sum \mathbf{C}_i \boldsymbol{b}_i)$$





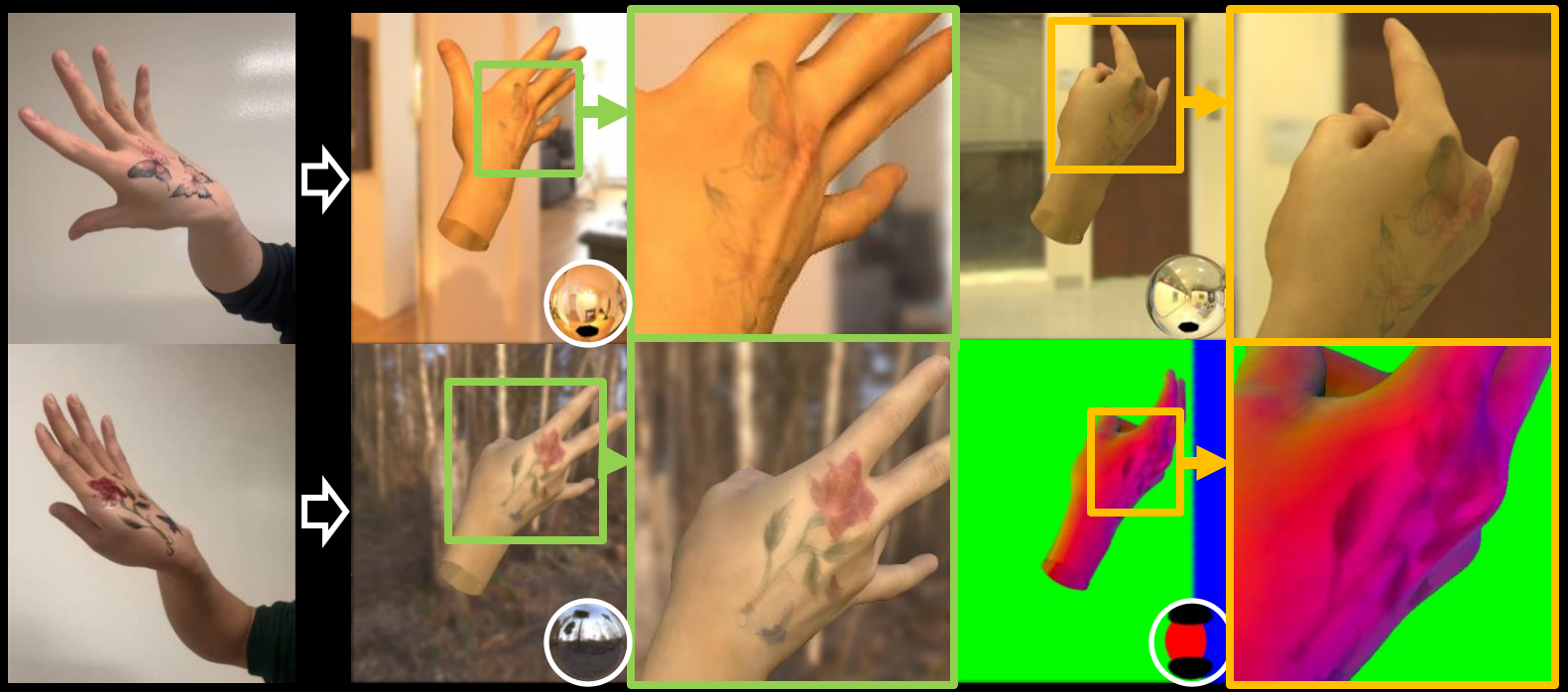
Non-Linear Network

Remove Bias

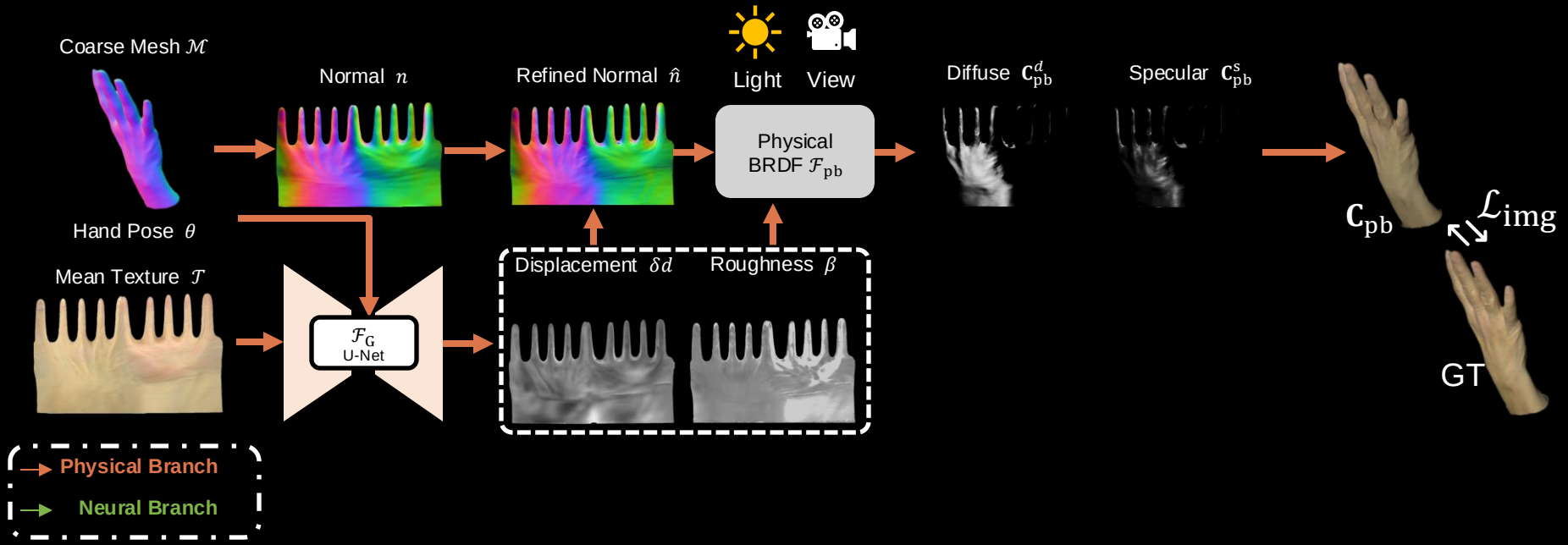
Remove Non-linear Activations

Spatially Varying Linear Network

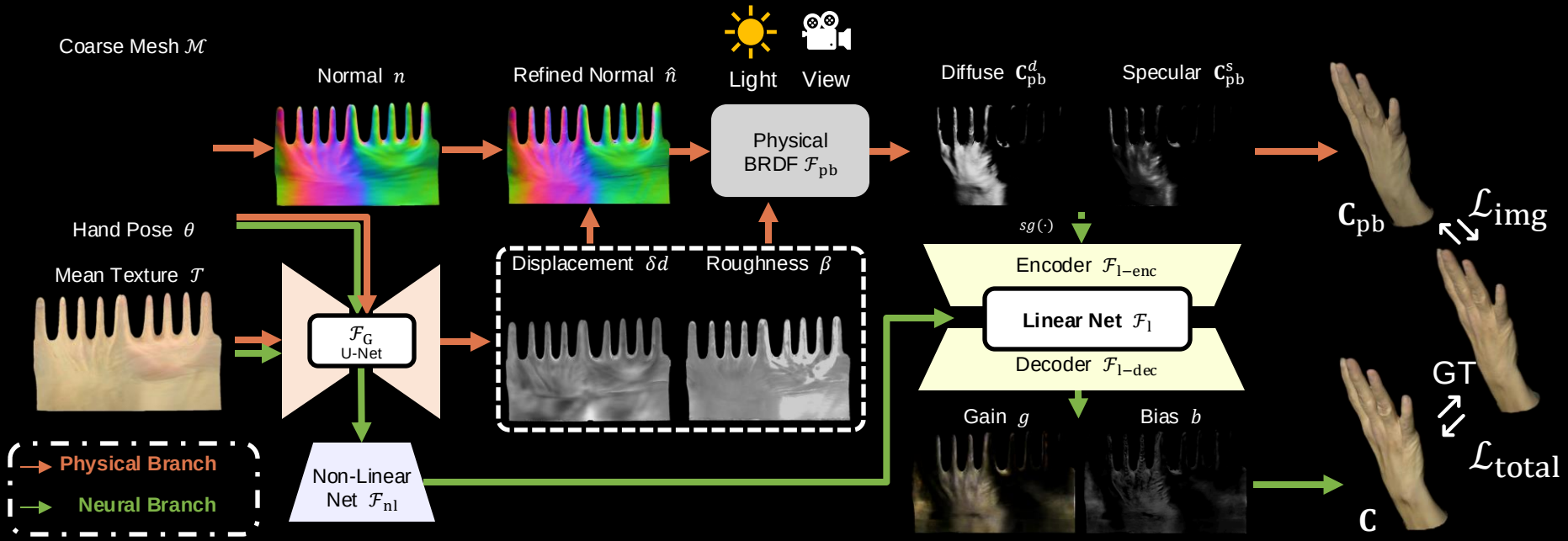
Key Idea: Linear Lighting Model



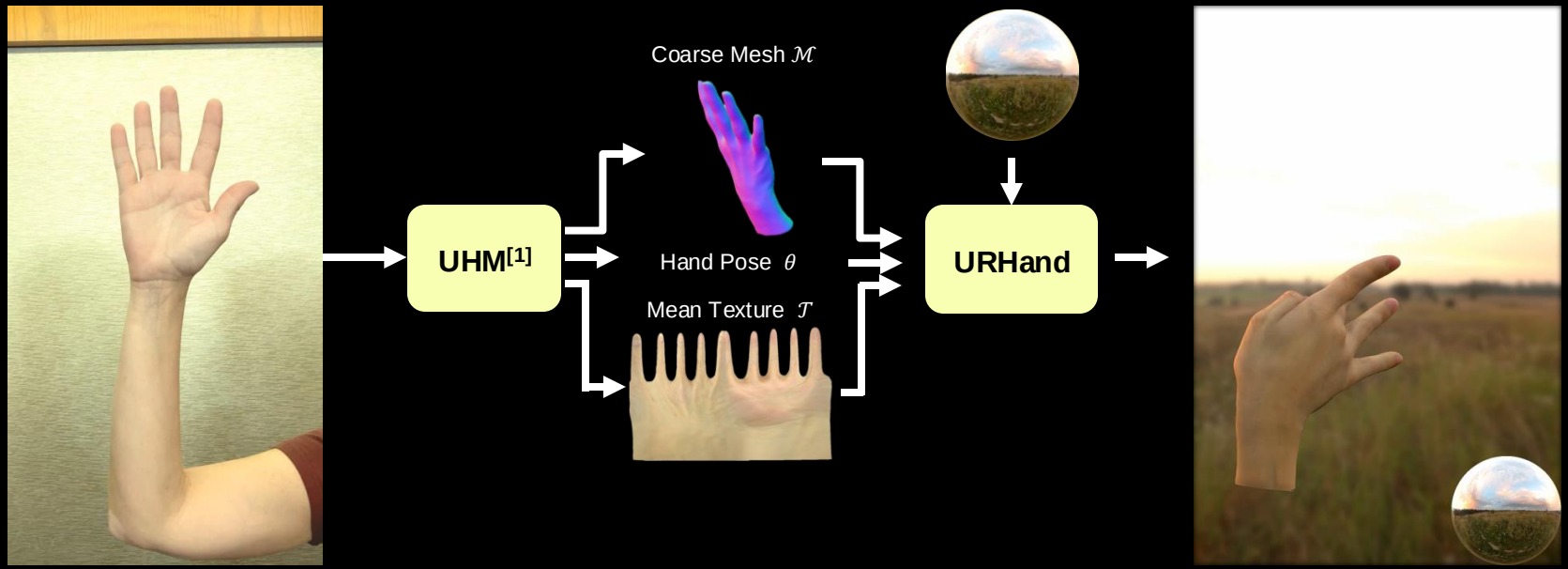
URHand: Physical Branch



URHand: Neural Branch



URHand: Inference



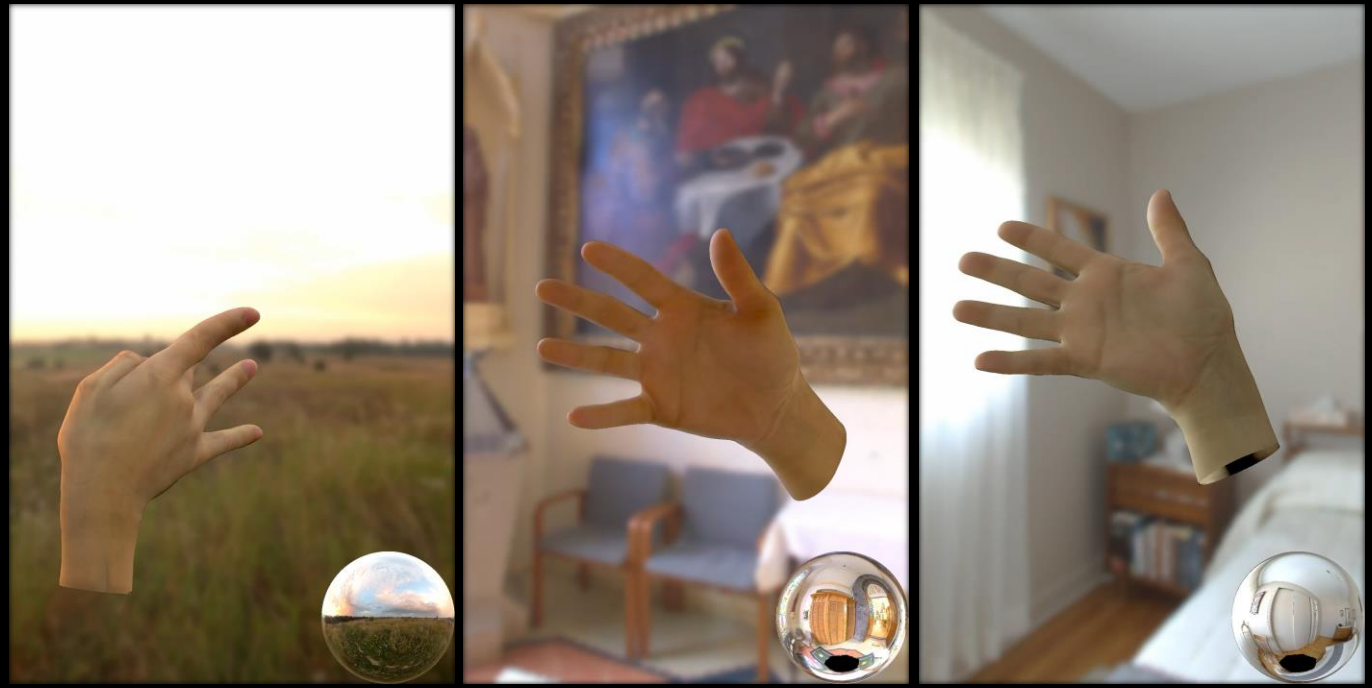
[1] Moon et al. “Authentic Hand Avatar from a Phone Scan via Universal Hand Model”. CVPR. 2024.

Relight with Environment Map

URHand ⇨ Your Hand

Quick Personalization of a Relightable Hand from Phone Scan

Input Phone Scan



Photorealistic Rendering under Arbitrary Illuminations

Expressive Whole-Body 3D Gaussian Avatar

https://mks0601.github.io/ExAvatar/

Gyeongsik Moon

Takaaki Shiratori

Shunsuke Saito

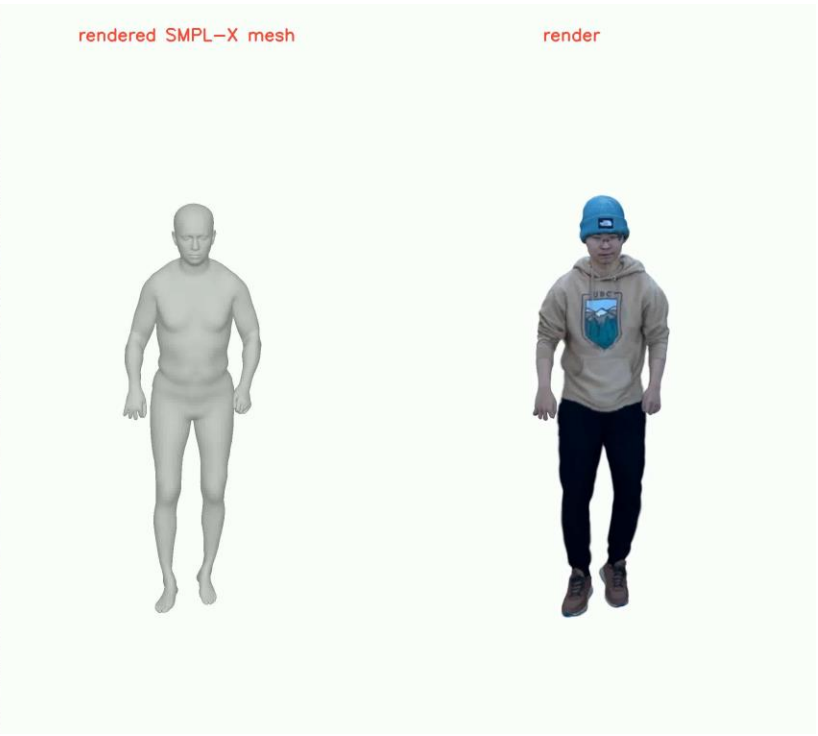
DGIST
Daegu Gyeongbuk Institute of Science & Technology

Meta

Expressive *Whole-Body* 3D Avatar from a *Monocular Video*

A monocular video

Driving with novel *body poses, hand poses, and facial expressions*



[1] Moon et al. "Expressive Whole-Body 3D Gaussian Avatar". ECCV. 2024.
3D pose tracking is done with [2] Moon et al. "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation." CVPRW. 2022. Music is from Jungkook (BTS) – Standing Next To You. Dance is from Youtuber Vincent Hsu.

What makes it challenging?

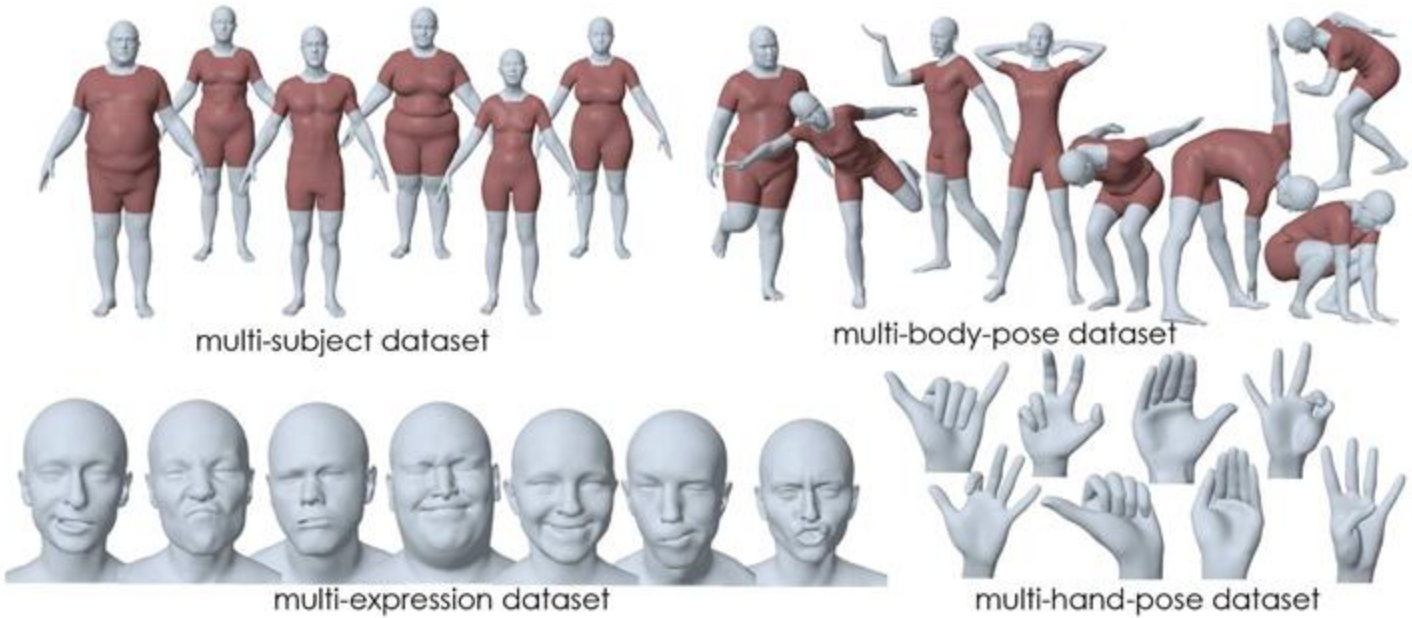
Limited training frames

- Limited body poses
- Limited hand poses
- Limited facial expressions

How can we make it generalize well to novel poses and facial expressions?

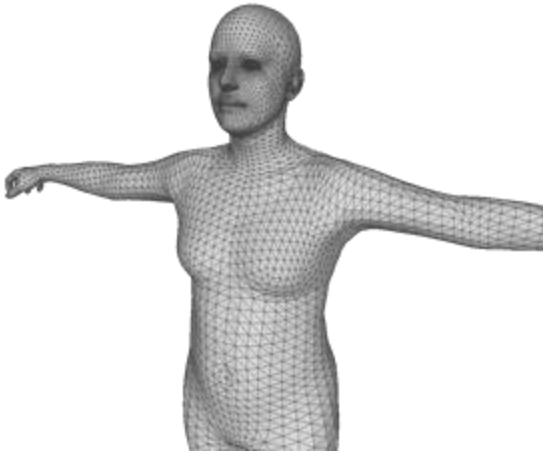
Strong Priors from Mesh-Based Models

- 3D mesh-based models (e.g., SMPL-X [1] and GHUM [2]) already support animation with novel body poses, hand poses, and facial expressions
- They provide strong geometric priors for the 3D human animations

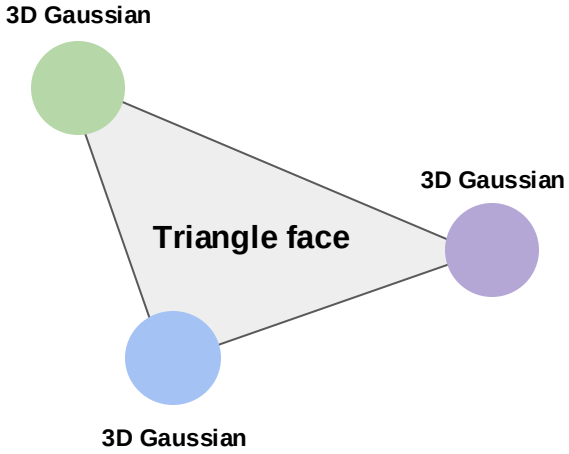


[1] Pavlakos et al. "Expressive body capture: 3D hands, face, and body from a single image." CVPR. 2019.
[2] Xu et al. "Ghum & ghuml: Generative 3D human shape and articulated pose models." CVPR. 2020.

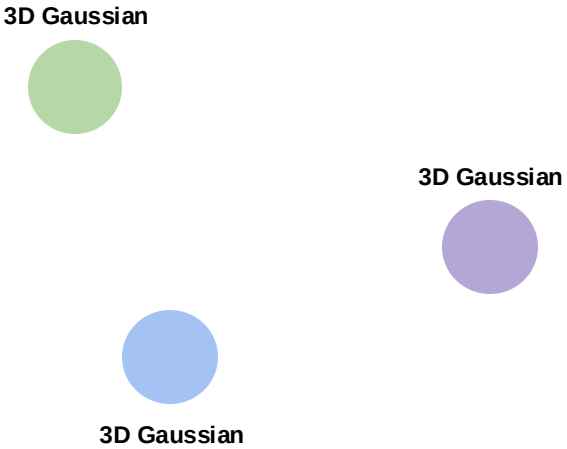
Hybrid Representation of 3DGS and Surface Mesh



Surface mesh
+ geometric priors
- hard to model clothes and hair



Hybrid (Ours)
+ geometric priors
+ can model clothes and hair



3D Gaussians
+ can model clothes and hair
- no geometric priors

Hybrid Representation of 3DGS and Surface Mesh

We can utilize useful surface-based regularizers
thanks to the hybrid representation!

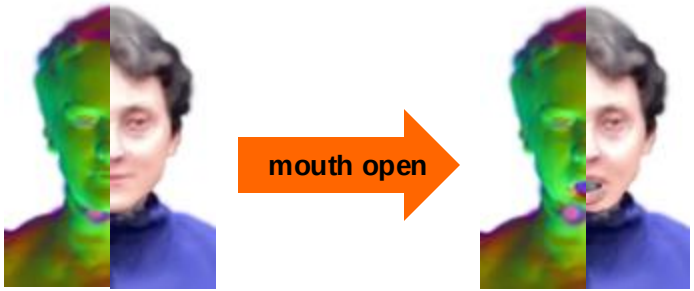
(a) With Lap. reg. (Ours)

(b) Without Lap. reg.

(c) Without Lap. reg. + strong L2 reg.

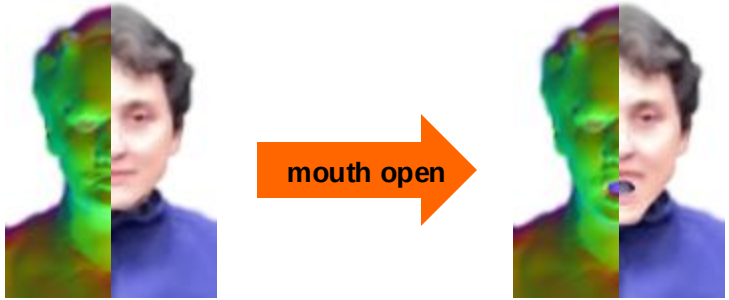
Hybrid Representation of 3DGS and Surface Mesh

We can utilize useful surface-based face loss
thanks to the hybrid representation!



(a) With the face loss (Ours)

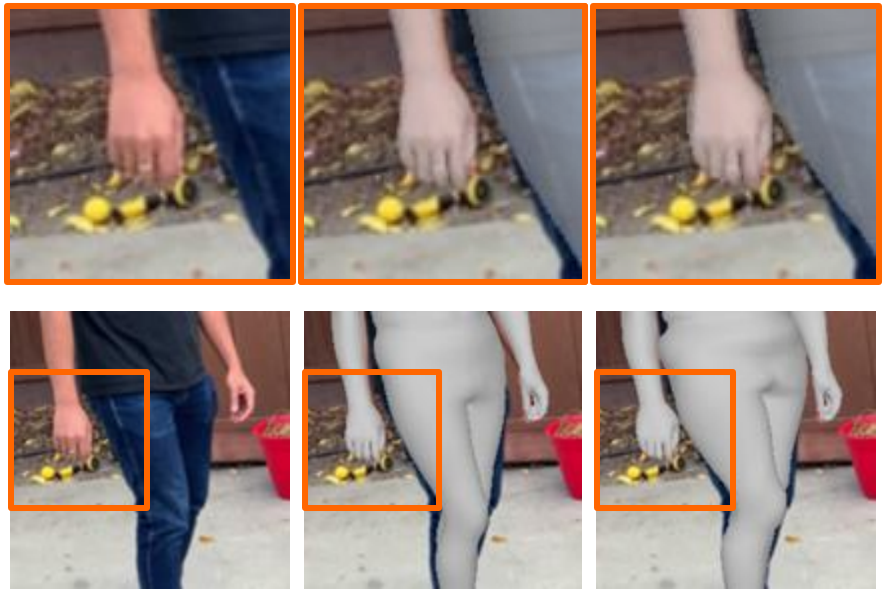
Mouth geometry is at the correct position



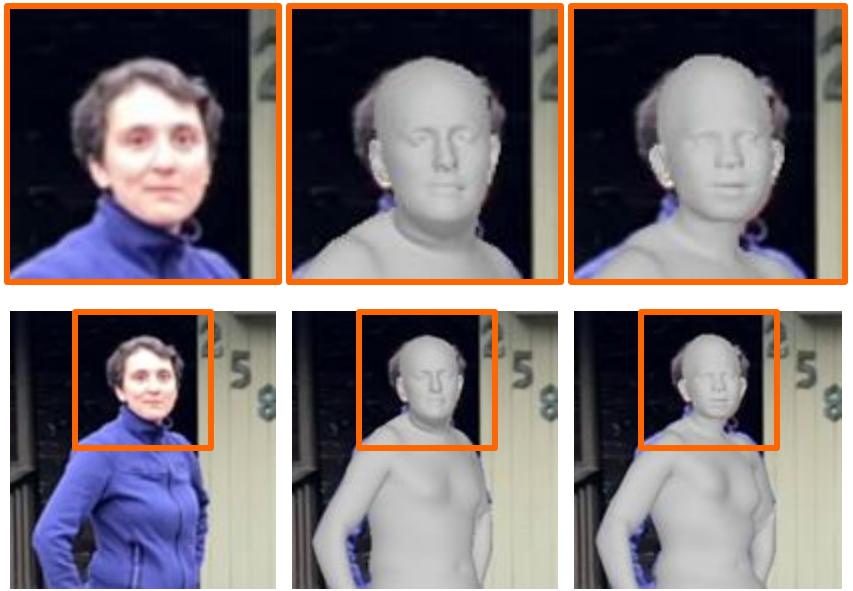
(b) Without the face loss

Mouth geometry is below the lower lip

Co-Registration of Body, Hands, and Face

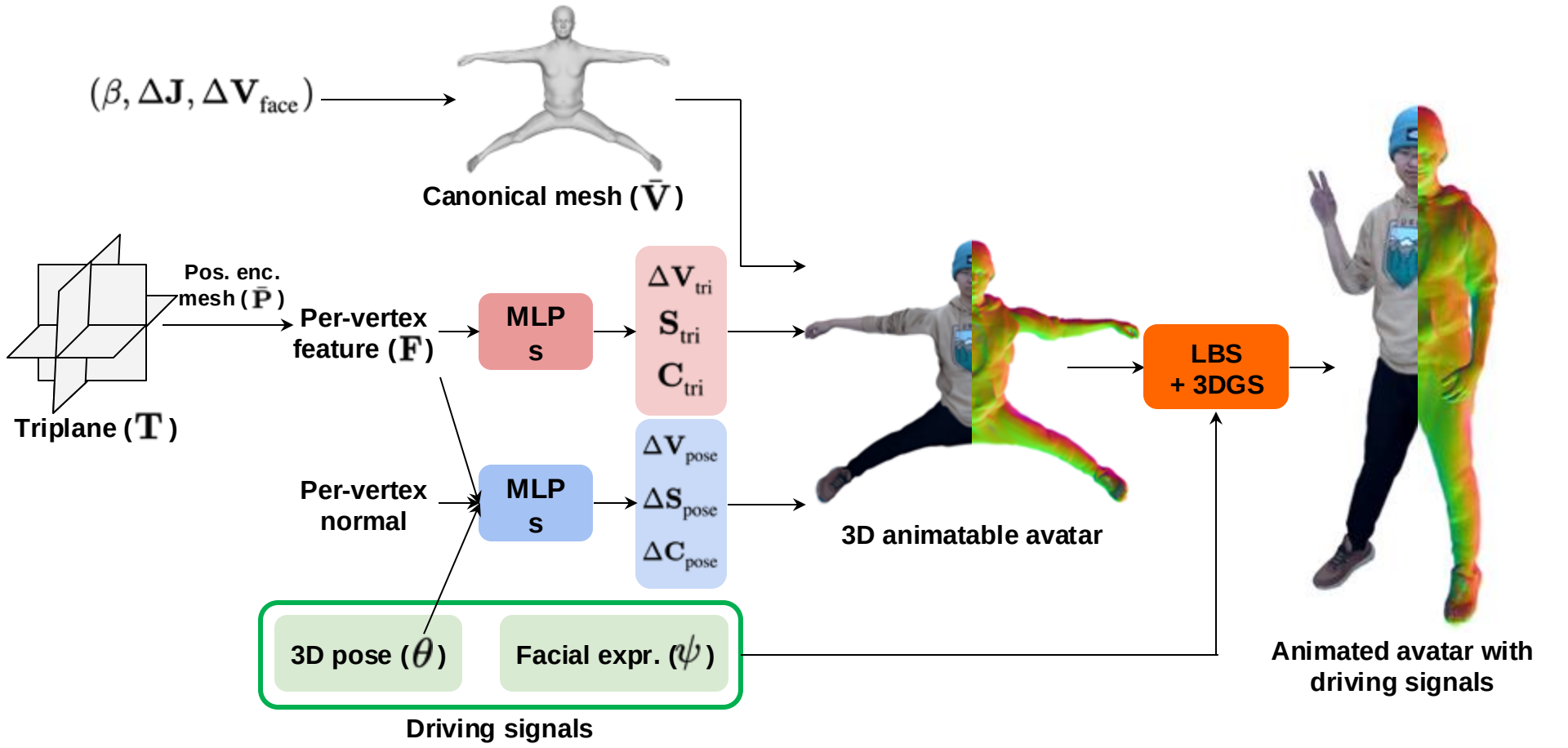


(a) Image | (b) With joint offset (Ours) | (c) Without joint offset



(a) Image | (b) With face offset (Ours) | (c) Without face offset

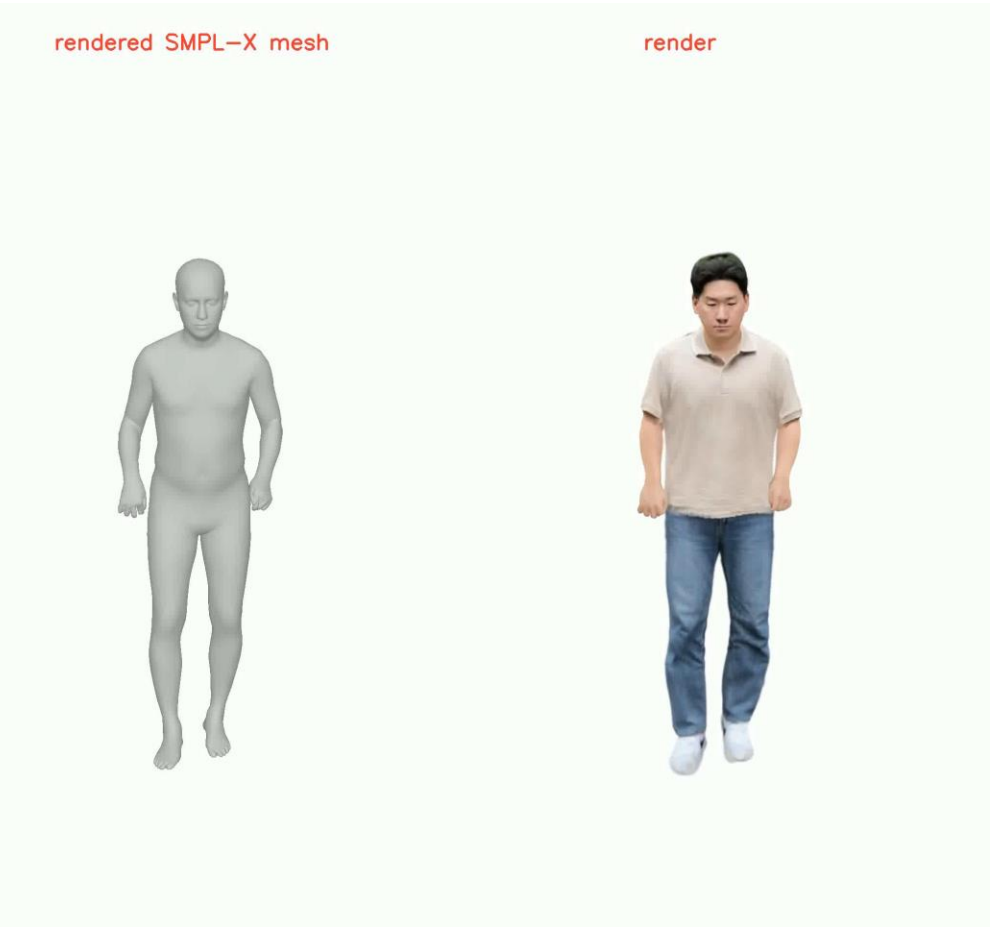
Architecture



Comparison to previous works

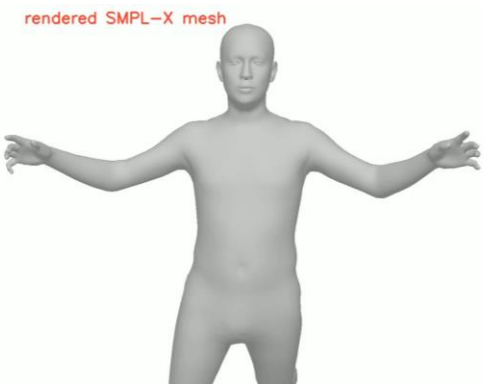
[1] Qian et al. "3DGS-Avatar: Animatable avatars via deformable 3D gaussian splatting." CVPR. 2024.

Motion transfer from in-the-wild videos



3D pose tracking is done with [1] Moon et al. "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation." CVPRW. 2022. Music is from Jungkook (BTS) – Standing Next To You. Dance is from Youtuber Vincent Hsu.

Motion transfer from in-the-wild videos

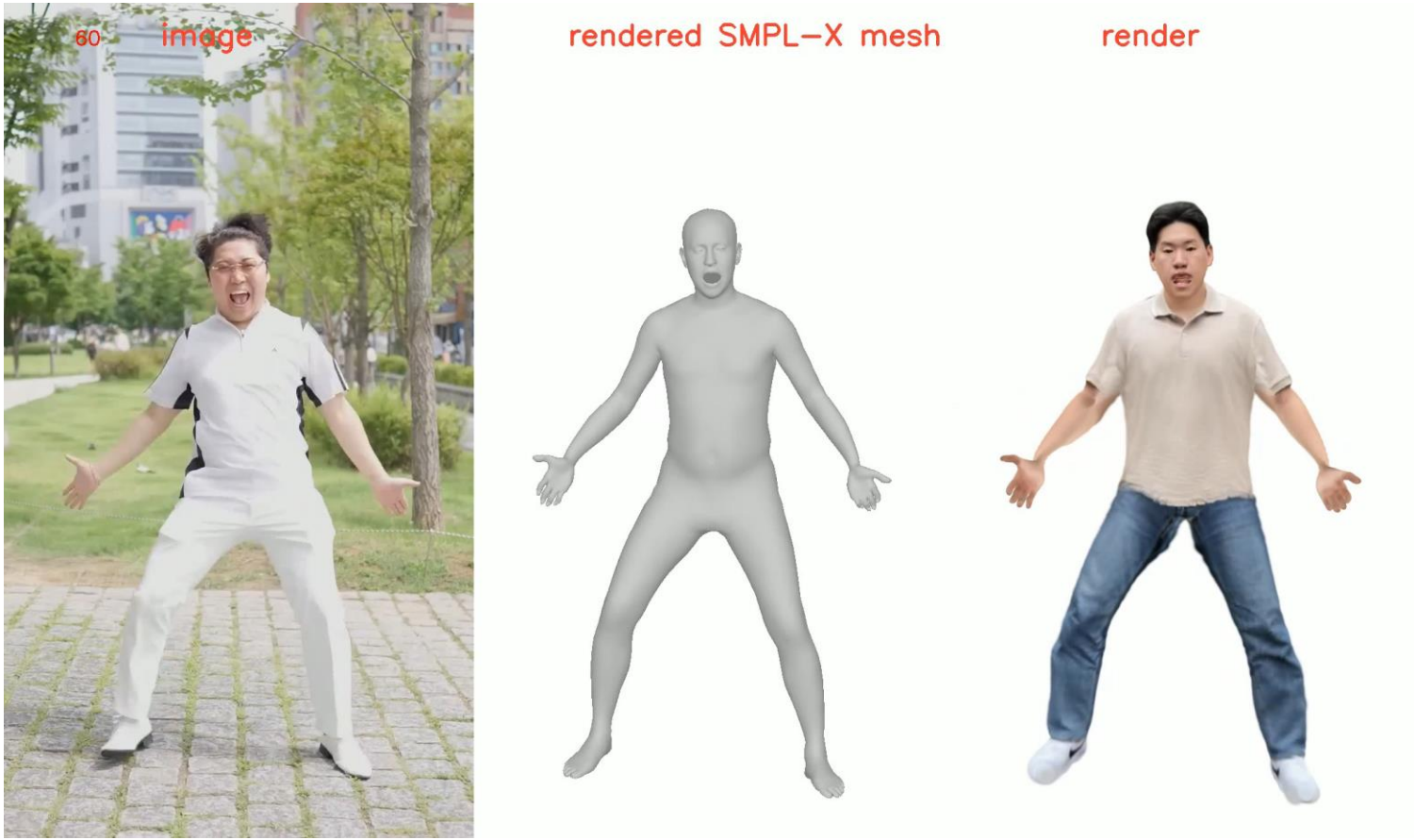


3D pose tracking is done with [1] Moon et al. "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation." CVPRW. 2022. Music is from Michael Jackson – Smooth Criminal. Dance is from YOOTAEYANG of SF9.

Motion transfer from in-the-wild videos

3D pose tracking is done with [1] Moon et al. "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation." CVPRW. 2022. Music and dance are from TVXQ – Mirotic.

Motion transfer from in-the-wild videos



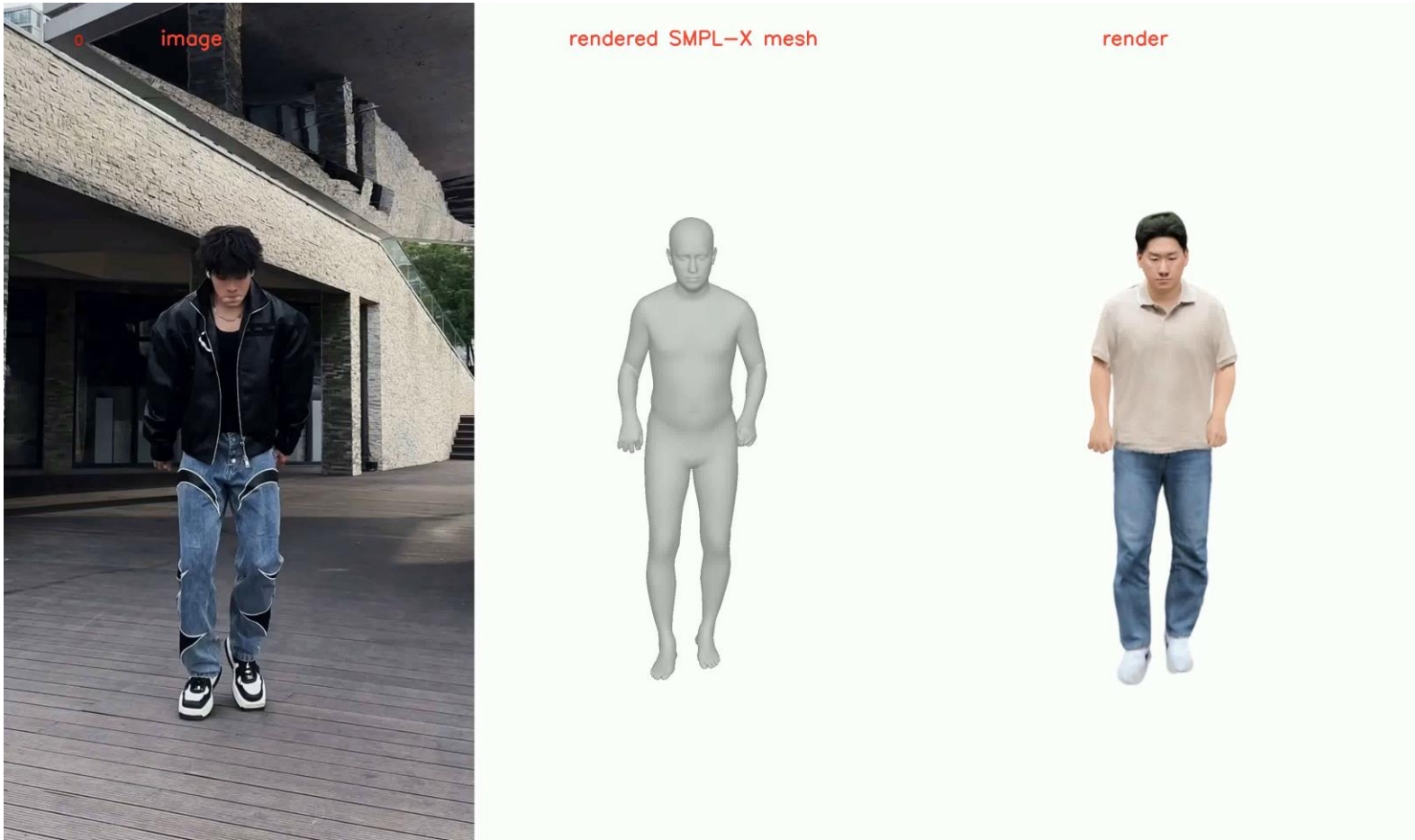
3D pose tracking is done with [1] Moon et al. "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation." CVPRW. 2022. Music and video are from Youtuber the myeonsang.

Motion transfer from in-the-wild videos

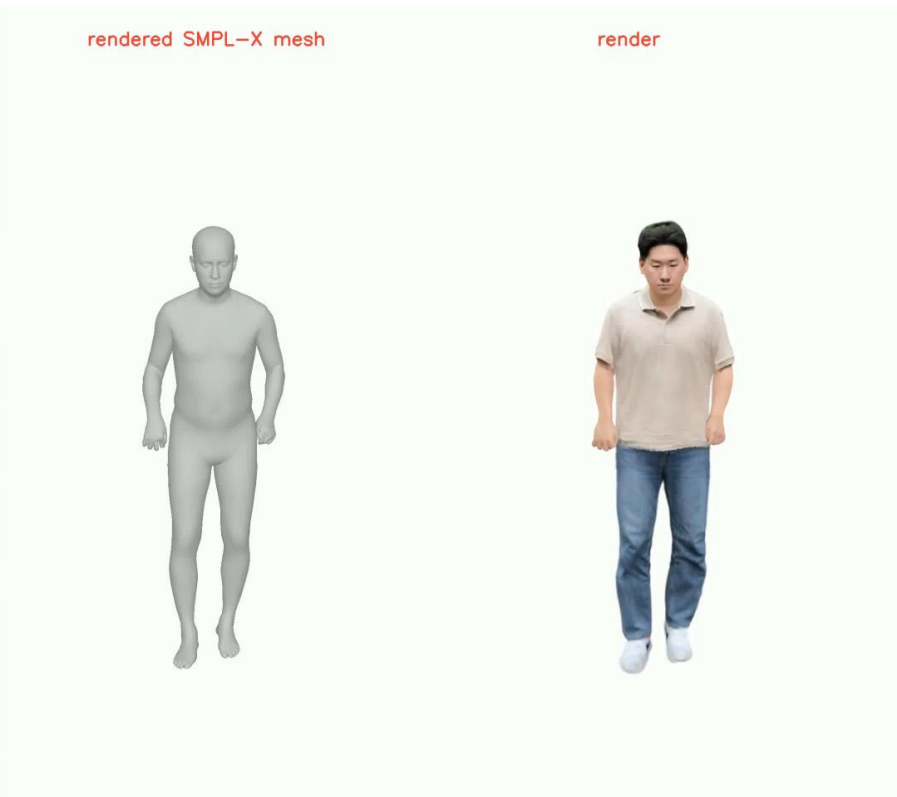
KOREA UNIVERSITY
Department of
Computer Science and Engineering

고려대학교 응원가 – 민족의 아리아

Render from any viewpoints with 3D avatar



We just need *a casually captured video*



3D pose tracking is done with [1] Moon et al. "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation." CVPRW. 2022. Music is from Jungkook (BTS) – Standing Next To You. Dance is from Youtuber Vincent Hsu.

Ph.D. DISSERTATION

Expressive Whole-Body 3D Multi-Person Pose and Shape Estimation from a Single Image

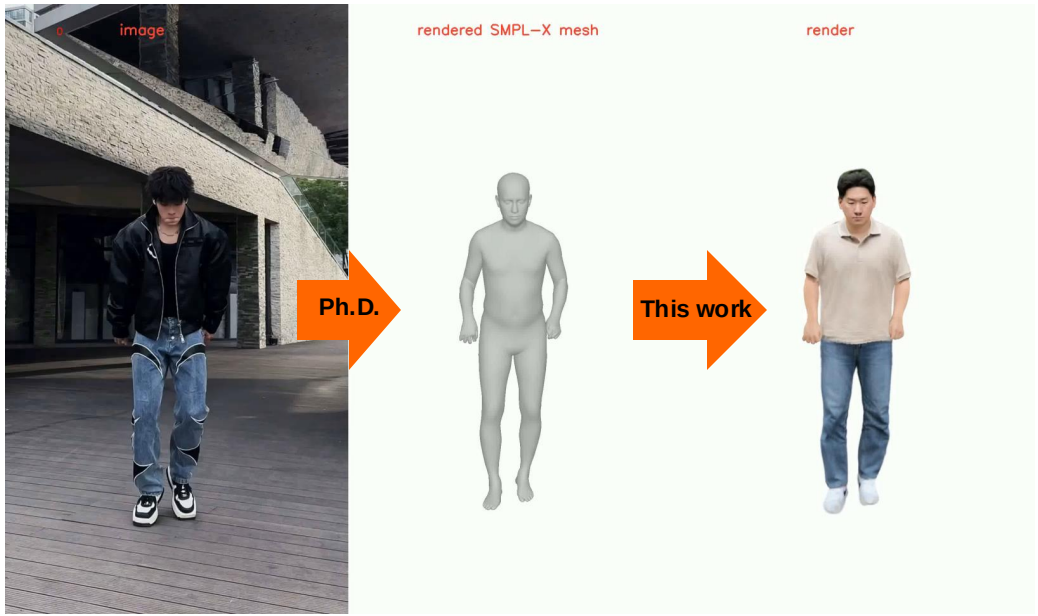
단일 이미지로부터 여러 사람의 표현적 전신 3D 자세 및 형태 추정

BY

GYEONGSIK MOON

February 2021

DEPARTMENT OF ELECTRICAL ENGINEERING AND
COMPUTER SCIENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY



PERSONA: Personalized Whole-Body 3D Avatar with Pose-Driven Deformations from a Single Image

Work in progress

How about a *single image*?

Artificially captured video

A single image is much more accessible than artificially captured videos

A single image

How about a *single image*?

Directly applying existing avatar works -> bad results due to limited visibility

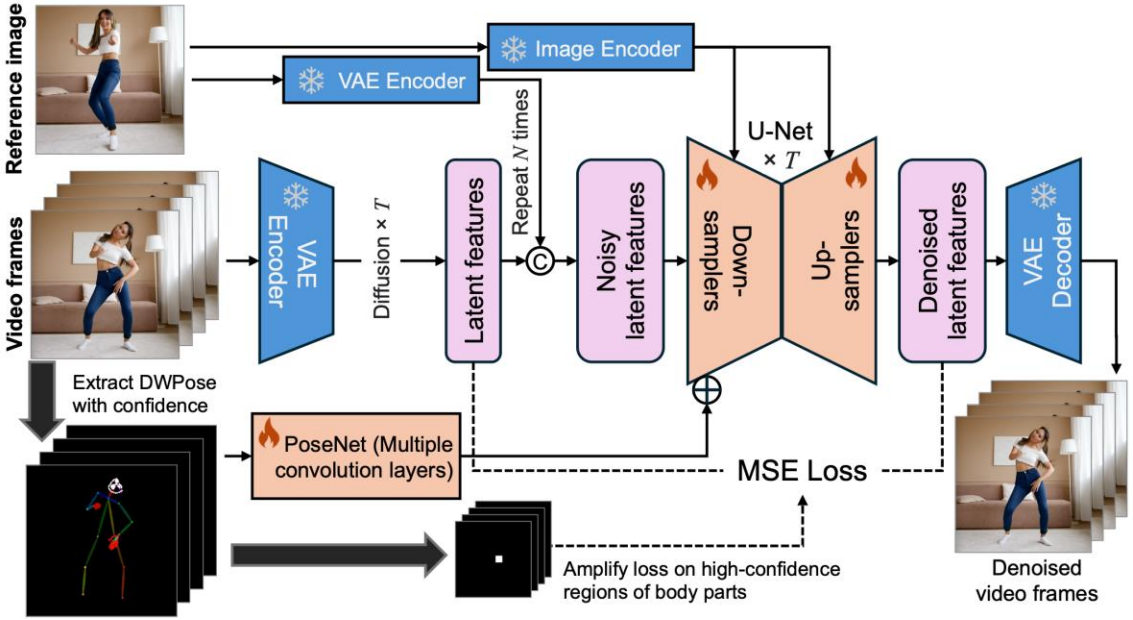
(a) Reference image

(b) Target pose

(c) ExAvatar

Prior from generative models

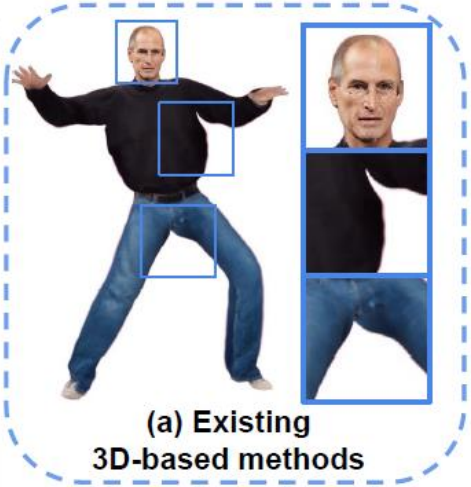
MagicAnimate [1]/AnimateAnyone [2]/MimicMotion [3]:
a person in a single reference image + 2D motion -> animated video



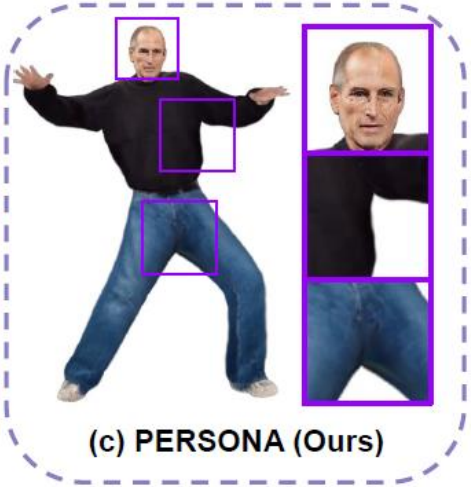
[1] Xu, Zhongcong, et al. "Magicanimate: Temporally consistent human image animation using diffusion model." CVPR. 2024.
[2] Hu, Li. "Animate anyone: Consistent and controllable image-to-video synthesis for character animation." CVPR. 2024.
[3] Zhang, Yuang, et al. "Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance." *arXiv preprint arXiv:2406.19680* (2024).

Input image

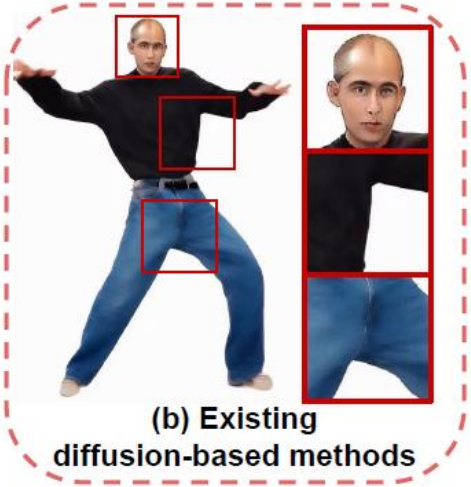
Authenticity



(a) Existing
3D-based methods



(c) PERSONA (Ours)

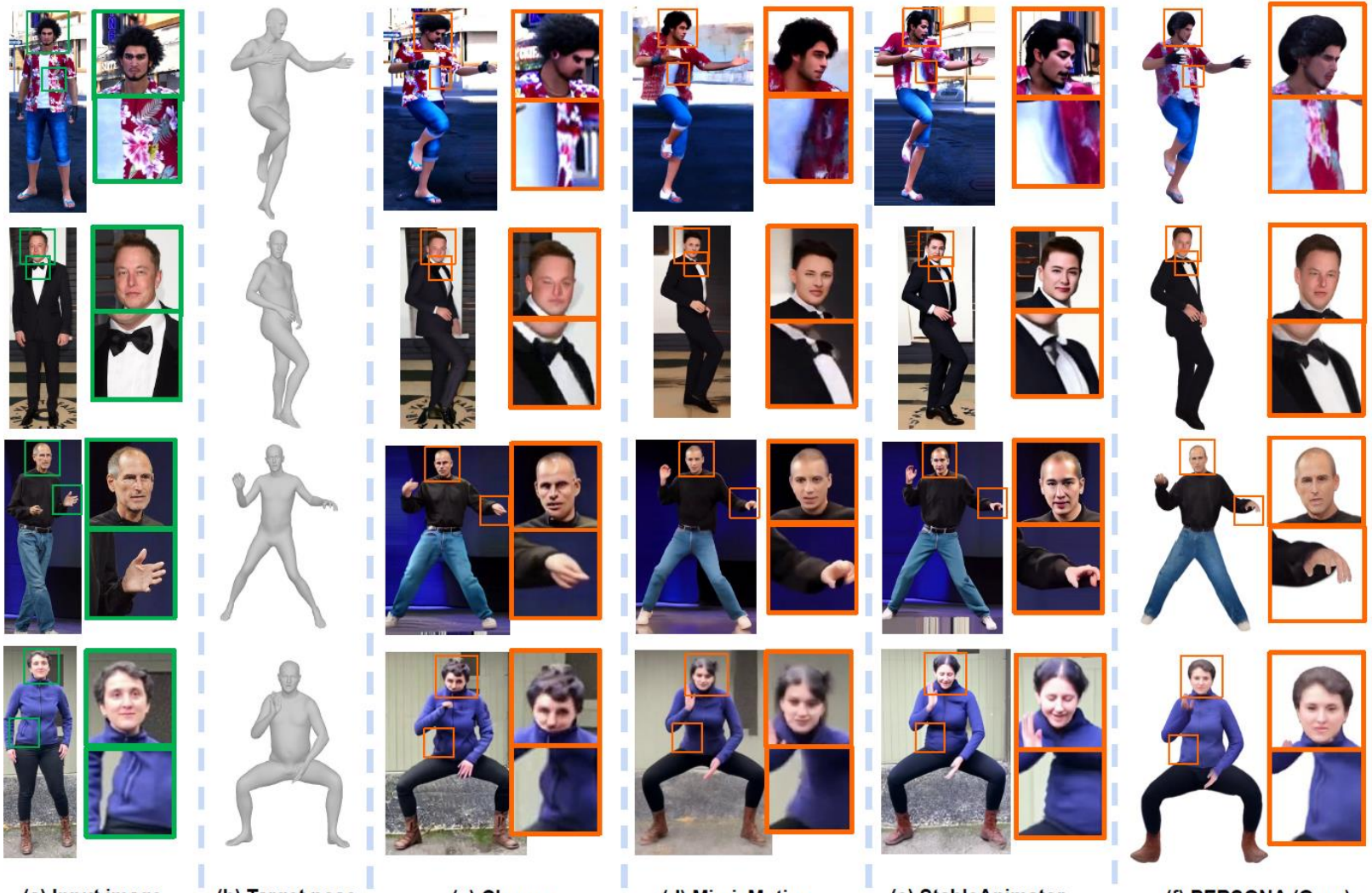


(b) Existing
diffusion-based methods

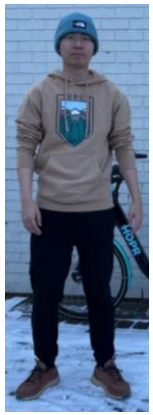
Pose-driven deformations


1375
Reference Image
Champ
MimicMotion
StableAnimator
AniGS
PERSONA (Ours)


Reference Image
Champ
MimicMotion
StableAnimator
AniGS
PERSONA (Ours)


(a) Input image
(b) Target pose
(c) Champ
(d) MimicMotion
(e) StableAnimator
(f) PERSONA (Ours)

Perceive and Represent Human Images



Perceive
(inverse rendering)

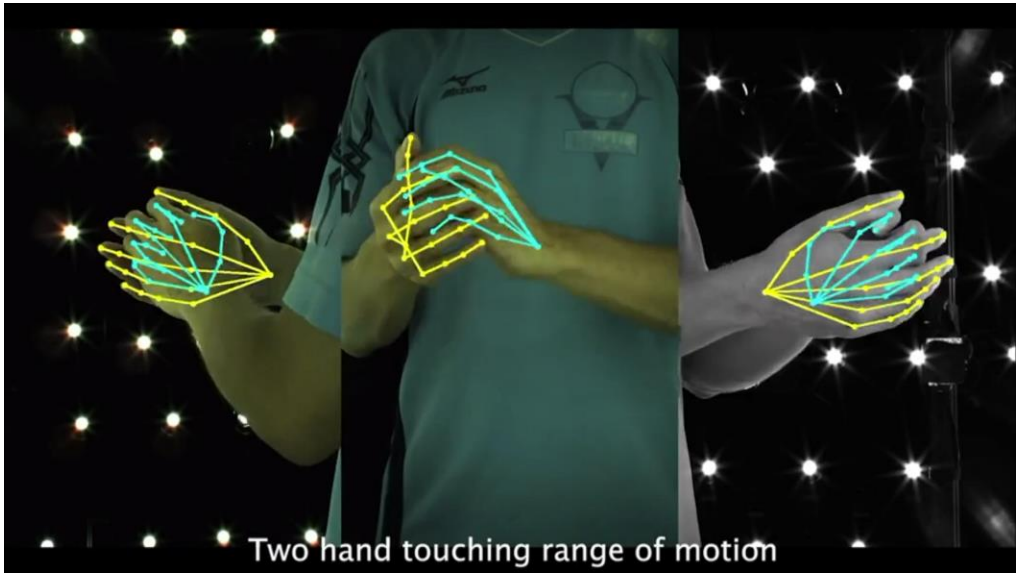
Control
(pose, light, …)

Represent
(rendering)

Human images

Human features
(e.g., geometry, texture, light, latent features, …)

3D Whole-Body Pose Estimation from a Single Image

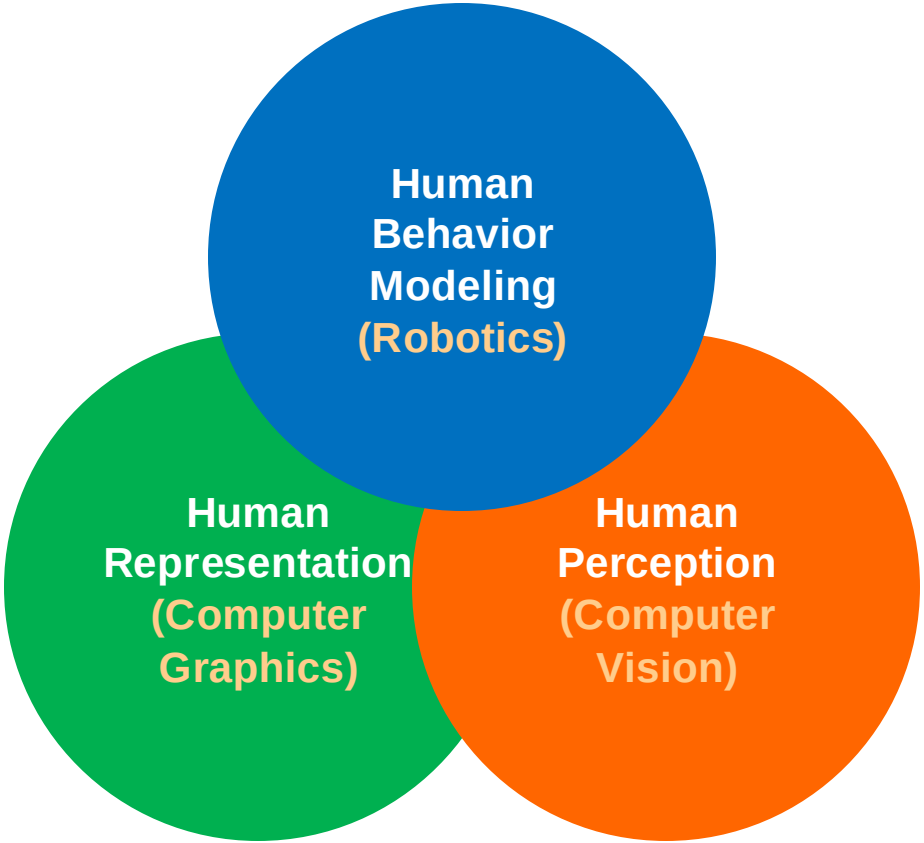


PARTE: Part-Guided Texturing for 3D Human Reconstruction from a Single Image

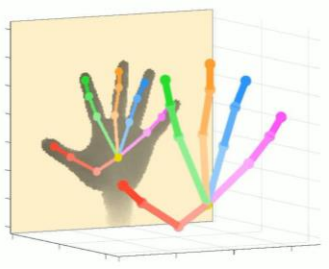
Work in progress

3D Whole-Body Pose + Surface + Texture Estimation from a Single Image

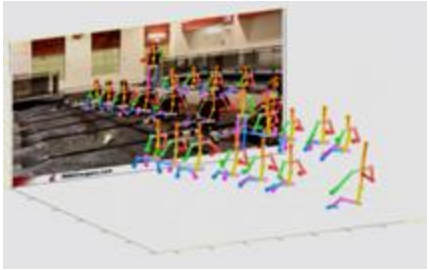
Visual Computing and AI Lab



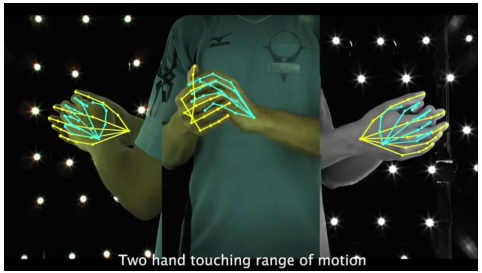
Computer Vision for Human Perception



3D hand pose estimation
(CVPR 2018)



3D multi-person pose estimation
(ICCV 2019)



3D interacting hand pose estimation
(ECCV 2020)

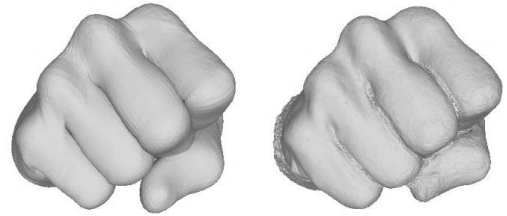
3D body shape estimation
(ECCV 2020)

3D body shape estimation
(CVPR 2021)

Hand-object interaction
(CVPR 2022)

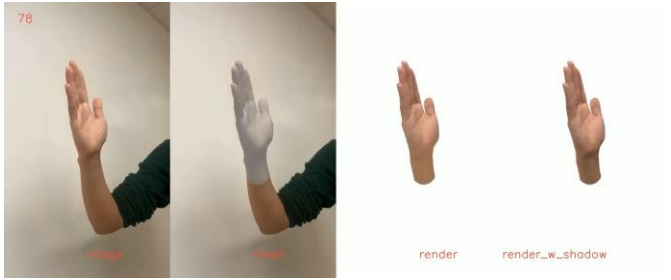
3D whole-body pose estimation
(CVPR 2022)

Computer Vision for Human Representation

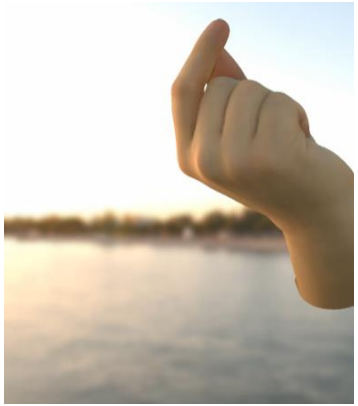


High-fidelity 3D hand geometry model
(ECCV 2020)

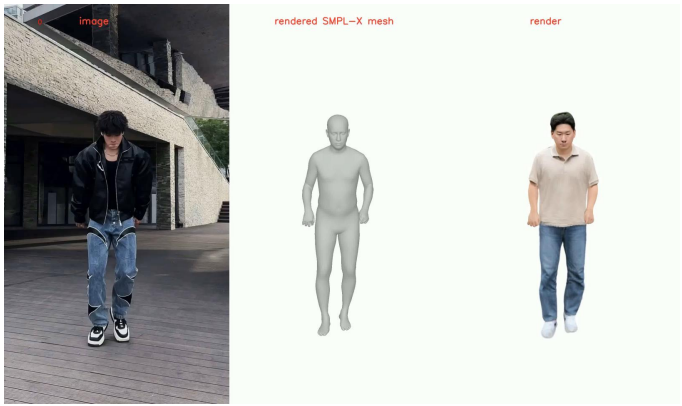
Relighted 3D interacting hands
(NeurIPS 2023)



Authentic 3D hand avatar
(CVPR 2024)

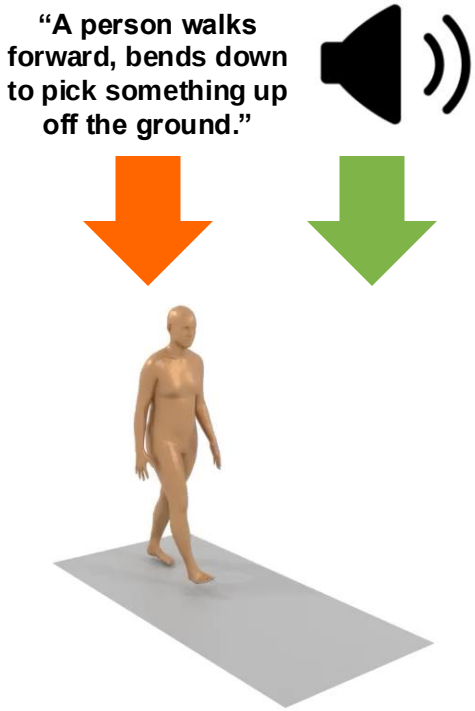


Universal relightable hand model
(CVPR 2024)



Expressive Whole-Body 3D Gaussian Avatar
(ECCV 2024)

Robotics for Human Behavior Modeling



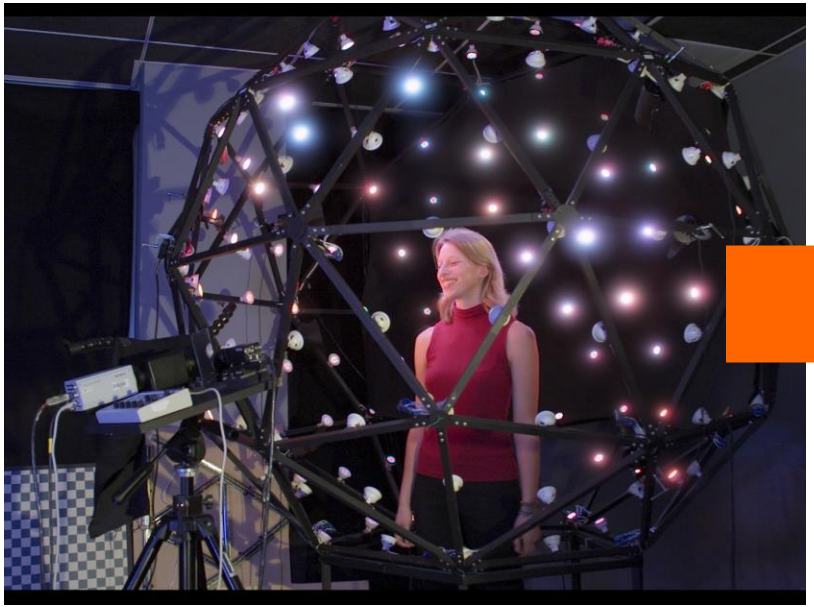
Human behavior modeling

Human-world interaction

Human-robot interaction

Understanding Humans in Casual Environments

- Casual environments have much less available observations
- Need to use **priors to complement limited observations**



Big Wave: Generative AIs

- They are so powerful without any human-specific priors (even without explicit concepts of 3D space)
- **Use priors from them**

OpenAI - SORA