Personalized 3D Human Avatar Generation and Real-Time Animation from a Single Image

Gyeongsik Moon

KOREA UNIVERSITY 1905
Visual Computing and AI Lab

1. CV+CG-based methods

- Previous works
- ExAvatar (ECCV 2024)

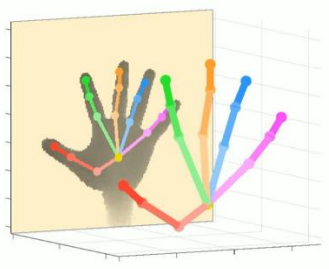
Break

2. CV+CG+GenAI-based methods

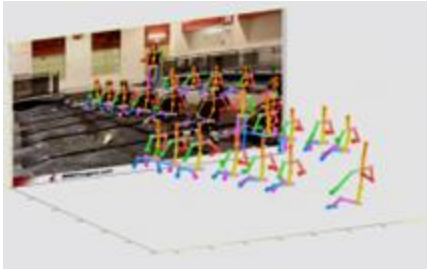
- Previous works
- PERSONA (ICCV 2025)

ComputerVisionLab
Seoul National University

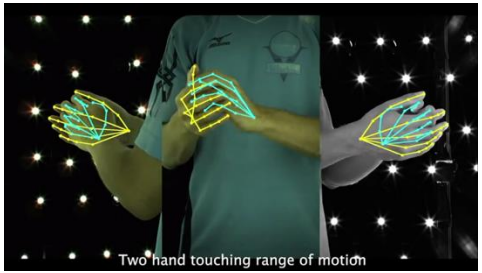
Computer Vision



3D hand pose estimation
(CVPR 2018)



3D multi-person pose estimation
(ICCV 2019)



3D interacting hand pose estimation
(ECCV 2020)

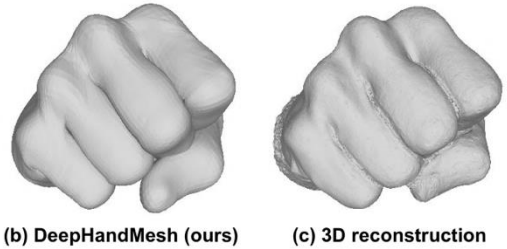
3D body shape estimation
(ECCV 2020)

3D body shape estimation
(CVPR 2021)

Hand-object interaction
(CVPR 2022)

3D whole-body pose estimation
(CVPR 2022)

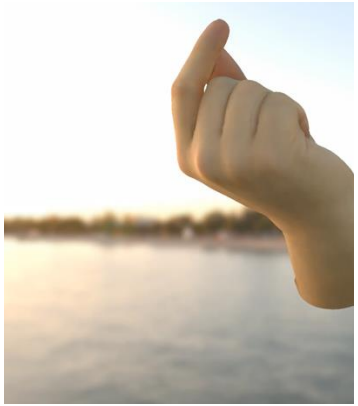
Computer Graphics



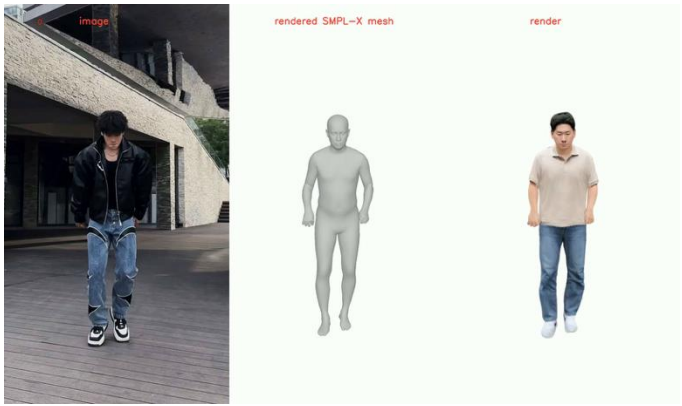
High-fidelity 3D hand geometry model
(ECCV 2020)

Relighted 3D interacting hands
(NeurIPS 2023)

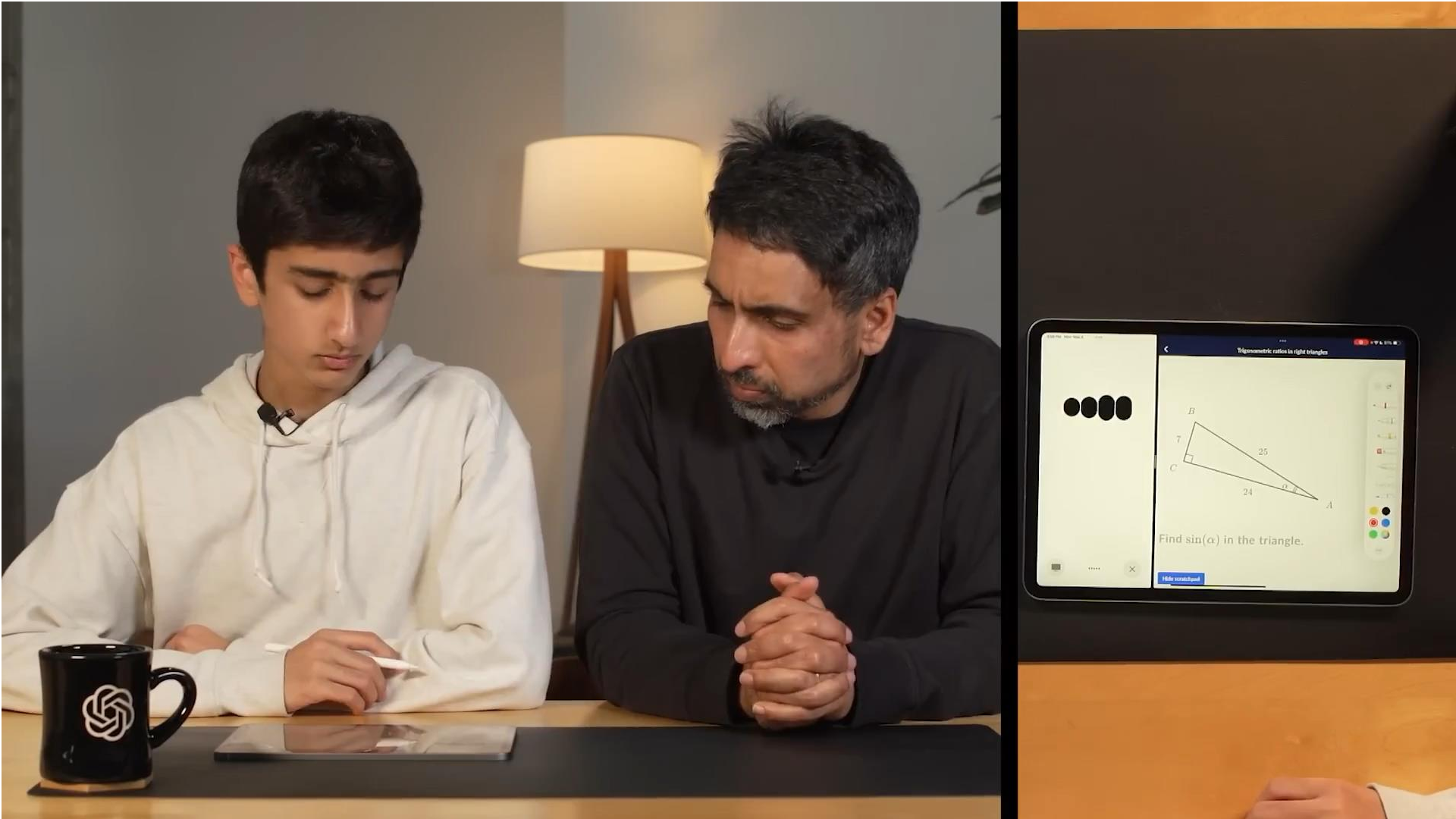
Authentic 3D hand avatar
(CVPR 2024)



Universal relightable hand model
(CVPR 2024)



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


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

Google Veo3 (I think it’s still good?)

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

>

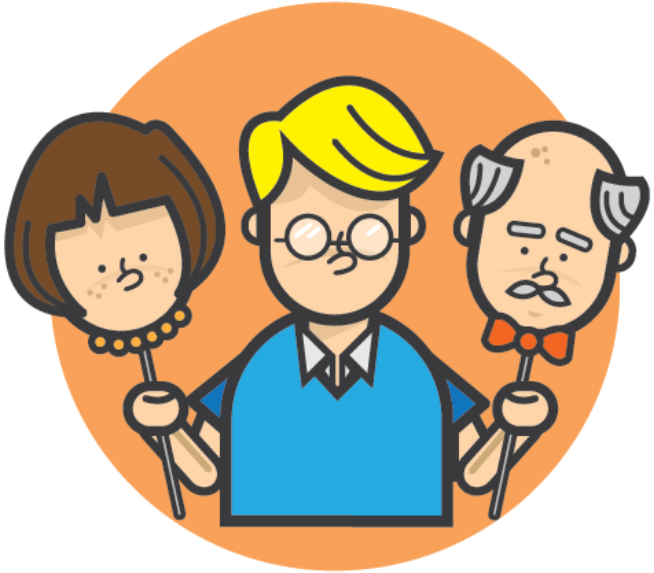
More challenging

Static scenes

per·so·na

/pərˈsōnə/

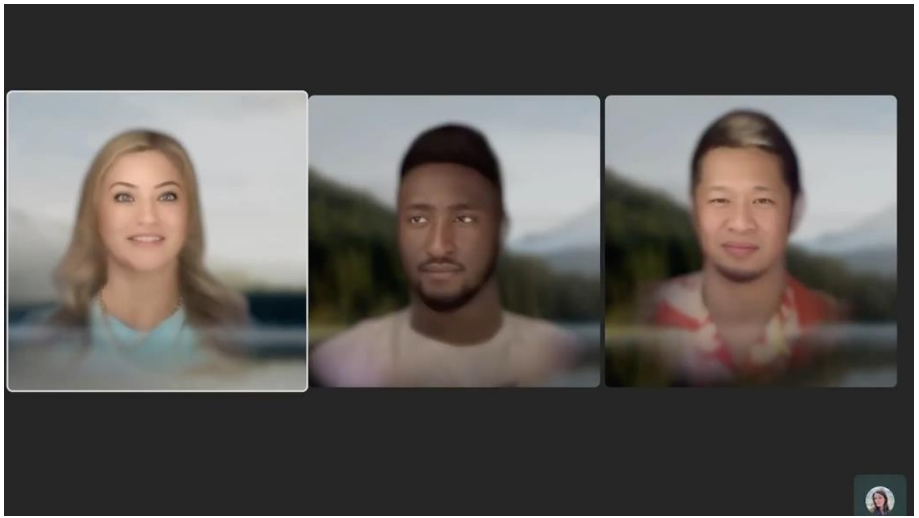
“a role or character adopted by an author, actor, etc. or in a game.”



Virtual characters in AAA movies and games

Telepresence

Codec Avatars (Meta)



Persona in Vision Pro (Apple)

Future: Interactive AI agents

LLM + Text-to-Speech + PERSONA

Interactive AI agents including non-verbal communications like humans

Animation with traditional pipeline

Manual 3D drawing
Character rigging in a canonical space

Marker-based MoCap.
Sync./calib. cameras and markers
Track actors' motion with them

Animated and rendered character

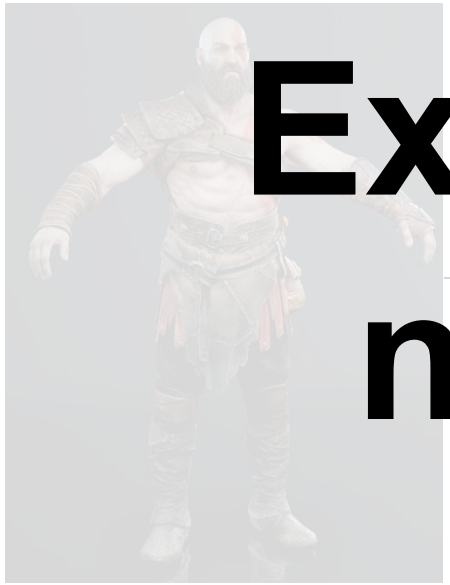
Renderer (animation, light, physics, …)

Animation with traditional pipeline

Manual 3D drawing
Character rigging in a canonical space

Marker-based MoCap.
Sync./calib. cameras and markers
Track actors' motion with them

Animated and rendered character



Renderer (animation, light, physics, …)

Expensive and not scalable

Persona from *Casual Inputs ((an) image(s))*

~~Manual 3D drawing~~

3D avatar creation

Character rigging in a canonical space

~~Marker-based MoCap.~~

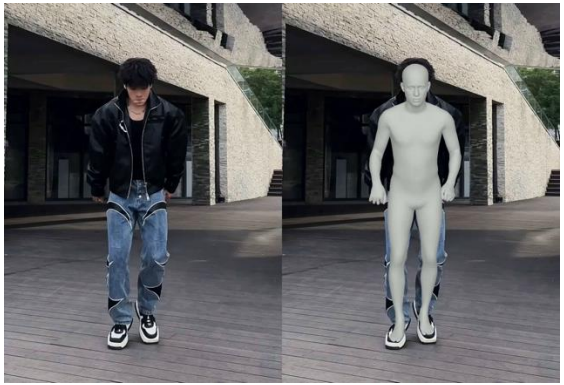
3D human pose est./gen.

Estimate/generate 3D poses in a canonical space

Animated and rendered character

Renderer (animation, light, physics, …)

Persona from *Casual Inputs ((an) image(s))*



~~Marker-based MoCap.~~

3D human pose est./gen.

Estimate/generate 3D poses in a canonical space

Animated and rendered character

Renderer (animation, light, physics, …)

Persona from *Casual Inputs ((an) image(s))*

~~Manual 3D drawing~~

3D avatar creation

Character rigging in a canonical space

Renderer (animation, light, physics, …)

Animated and rendered character

Early works: Neural Body (CVPR 2021. Best Paper Candidate.)
Training: Multi-view videos + accurate tracked 3D poses + NeRF
Animation: Any 3D poses

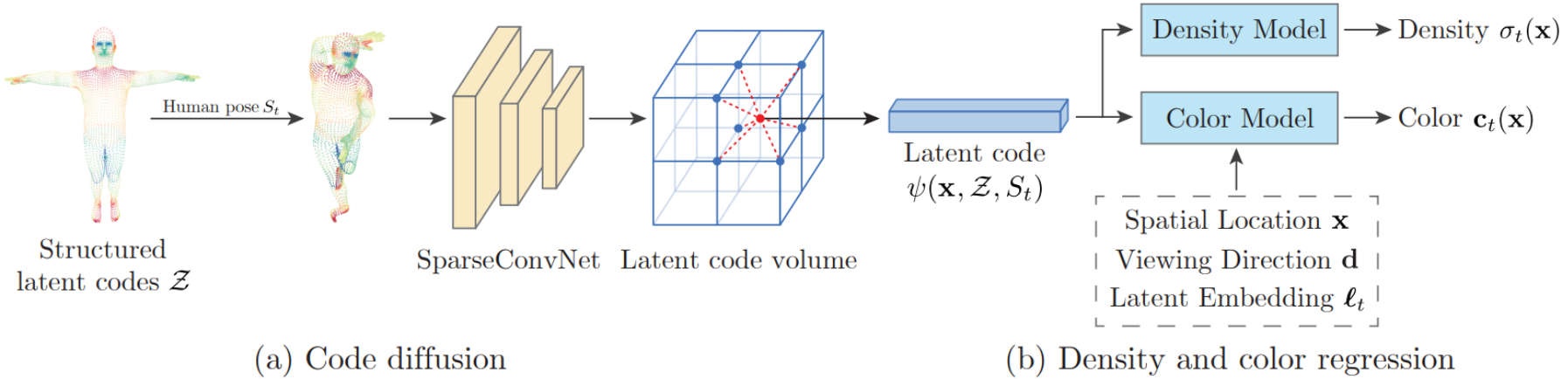
Task: Novel view synthesis from a sparse multi-view video

4-view video

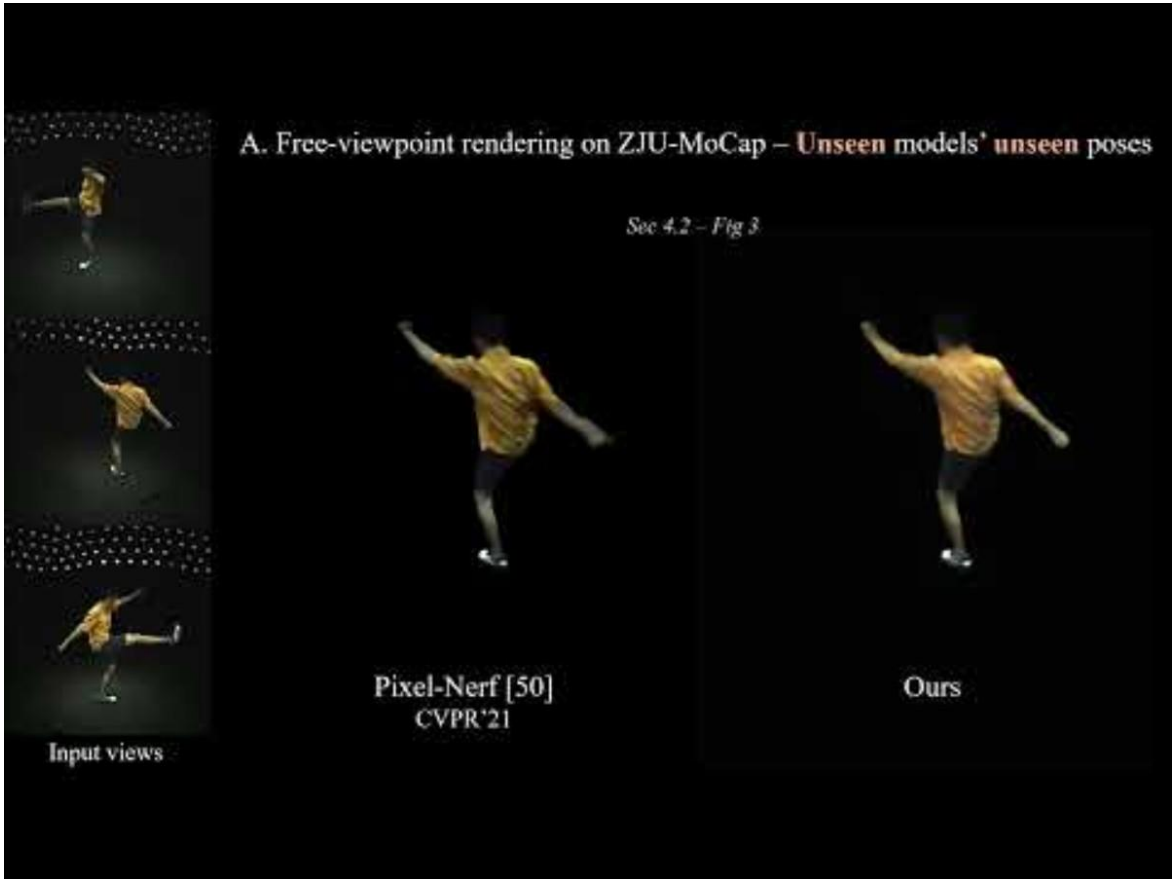
Novel view synthesis of dynamic human
(Our result)

Early works: Neural Body (CVPR 2021. Best Paper Candidate.)
Training: Multi-view videos + accurate tracked 3D poses + NeRF
Animation: Any 3D poses

- Build a human representation in canonical space (T-pose space)
  - Animate the canonical human with 3D pose+LBS
  - Use appearance models (NeRF or 3DGS) and render

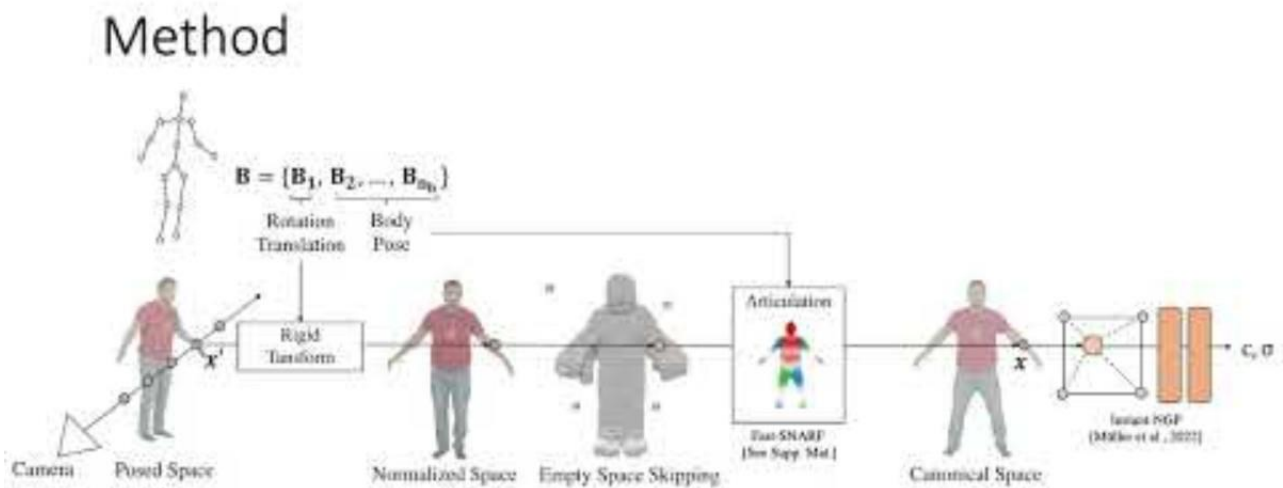


Early works: Neural Human Performer (NeurIPS 2021)
Training: Multi-view videos + accurate tracked 3D poses + NeRF
Animation: Any 3D poses



Early works: InstantAvatar (CVPR 2023)
Training: Monocular video + casually tracked 3D poses + NeRF
Animation: Any 3D poses

Now, moved from multi-view setup to casual setup!



We now have 3DGS!

This video contains a voice-over

3D Gaussian Splatting for Real-Time Radiance Field Rendering

SIGGRAPH 2023
(ACM Transactions on Graphics)

Bernhard Kerbl*
Inria UNIVERSITÉ CÔTE D'AZUR

Georgios Kopanas*
Inria UNIVERSITÉ CÔTE D'AZUR

Thomas Leimkühler
mpii

George Drettakis
Inria UNIVERSITÉ CÔTE D'AZUR

* Denotes equal contribution

GraphDeco

Early works: GaussianAvatar (CVPR 2024)
Training: Monocular video + casually tracked 3D poses + 3DGS
Animation: Any 3D poses

Moved from NeRF to 3DGS for the real-time rendering!
In CVPR 2024, more than 50 3DGS avatar papers are published…

GaussianAvatar:
Towards Realistic Human Avatar Modeling
from a Single Video via Animatable 3D Gaussians

Liangxiao Hu[1], Hongwen Zhang[2], Yuxiang Zhang[3], Boyao Zhou[3], Boning Liu[3],
Shengping Zhang[1], Liqiang Nie[1]
[1]Harbin Institute of Technology, [2]Beijing Normal University, [3]Tsinghua University

	Raw data	Processed data	Appearance models	Animation signal	Granularity
NeuralBody (CVPR 2021)	Multi-view videos	Accurate 3D poses	NeRF	Any 3D poses	Body
Neural Human Performer (NeurIPS 2021)	Multi-view videos	Accurate 3D poses	NeRF		
InstantAvatar (CVPR 2023)	Monocular video	Noisy 3D poses	NeRF		
GaussianAvatar (CVPR 2024)	Monocular video	Noisy 3D poses	3DGS		
ExAvatar (ECCV 2024)	Monocular video	Noisy 3D poses	3DGS		Body+hands+face

Expressive Whole-Body 3D Gaussian Avatar

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

Gyeongsik Moon

Takaaki Shiratori

Shunsuke Saito

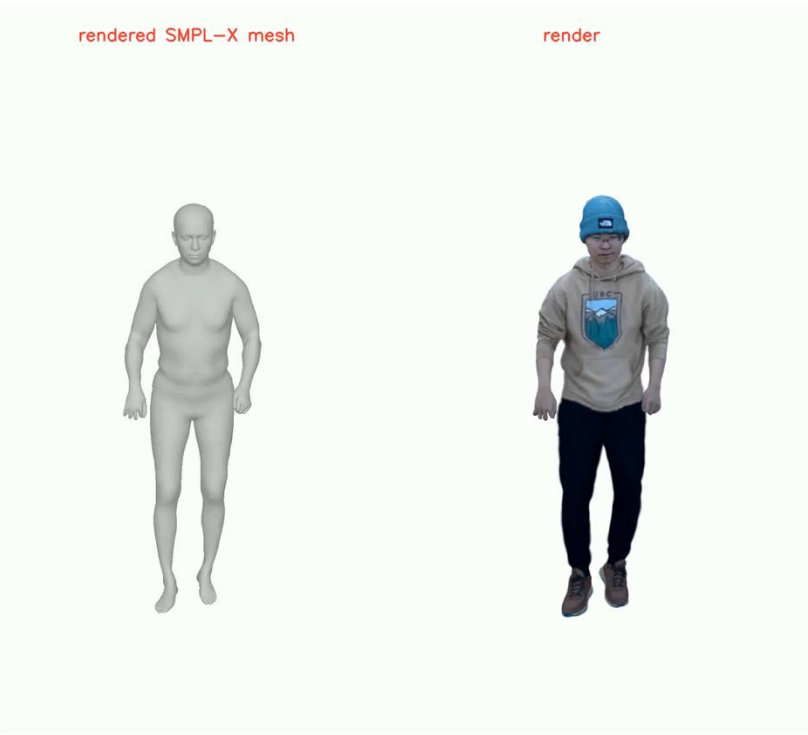
DGIST
Daegu Gyeongbuk Institute of Science & Technology

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?

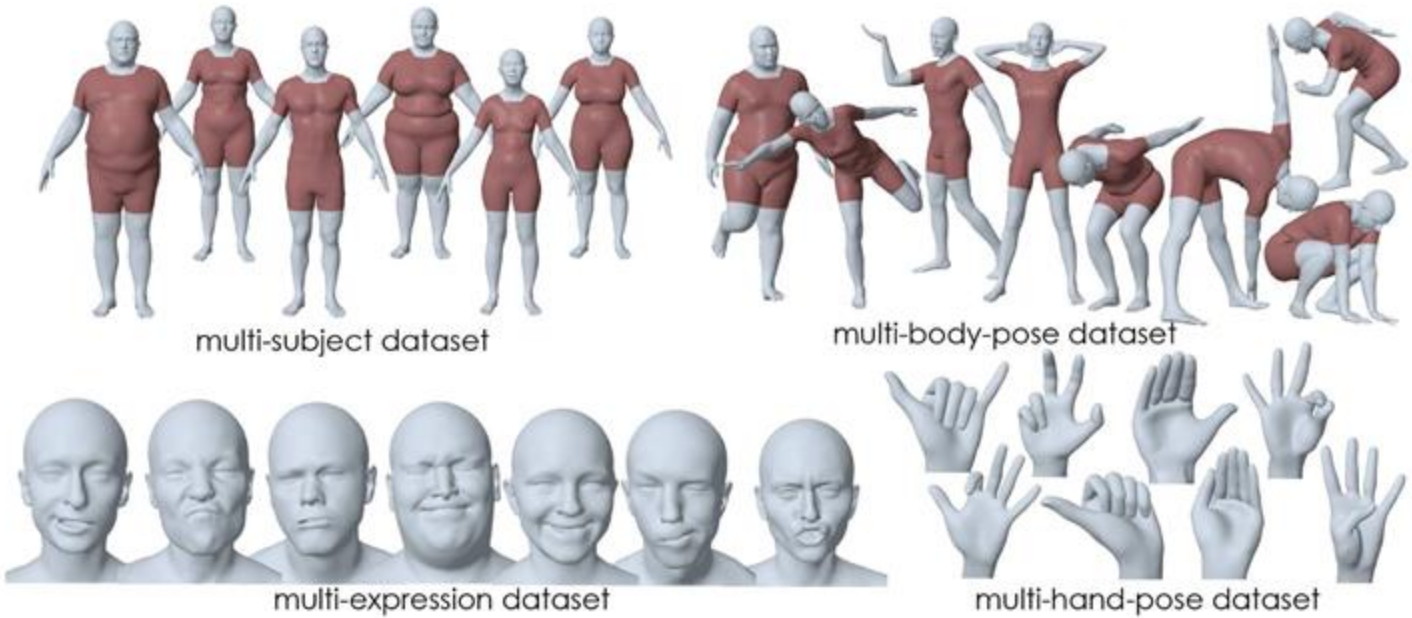
Limited training frames

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

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

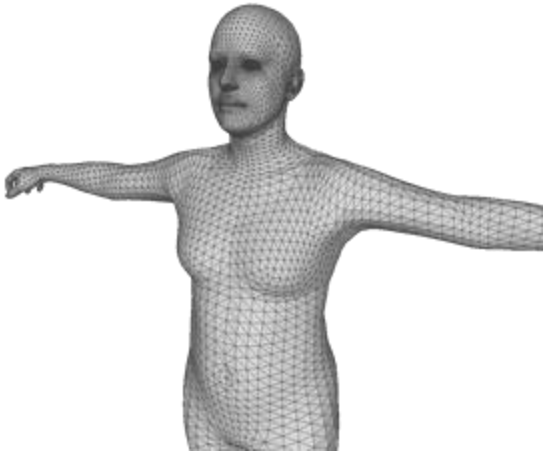
Strong Priors from Mesh-Based Models

- 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

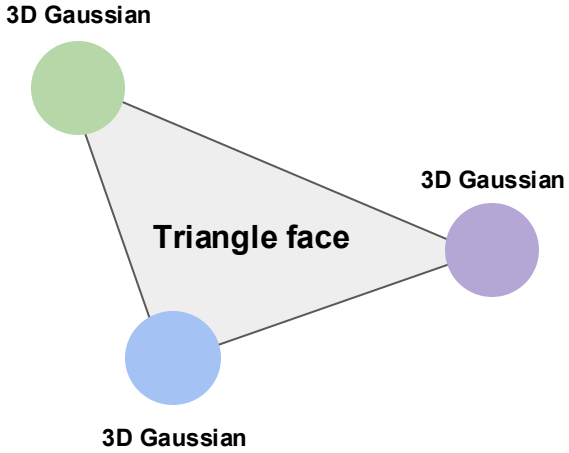


[1] Pavlakos et al. "Expressive body capture: 3D hands, face, and body from a single image." CVPR. 2019.
[2] Xu et al. "Ghum & ghuml: Generative 3D human shape and articulated pose models." CVPR. 2020.

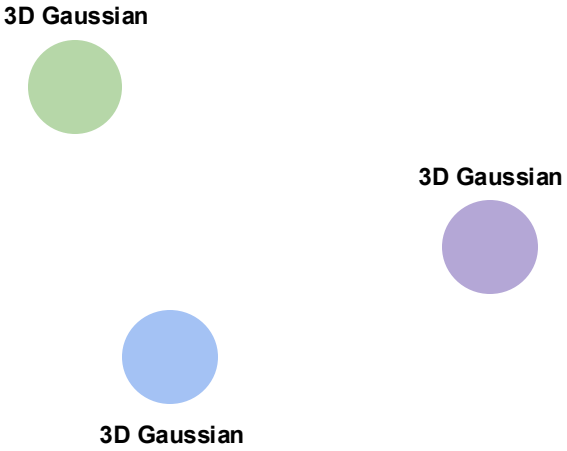
Hybrid Representation of 3DGS and Surface Mesh



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!

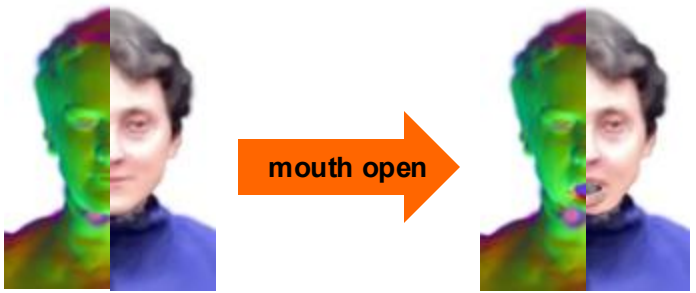
(a) With Lap. reg. (Ours)

(b) Without Lap. reg.

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

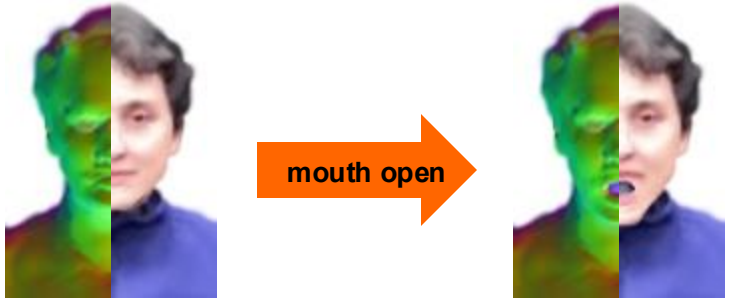
Hybrid Representation of 3DGS and Surface Mesh

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



(a) With the face loss (Ours)

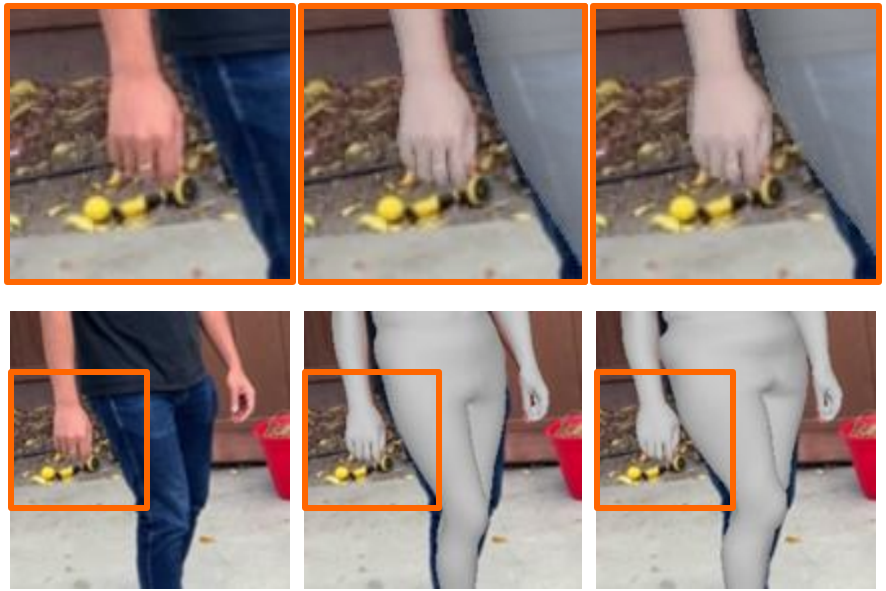
Mouth geometry is at the correct position



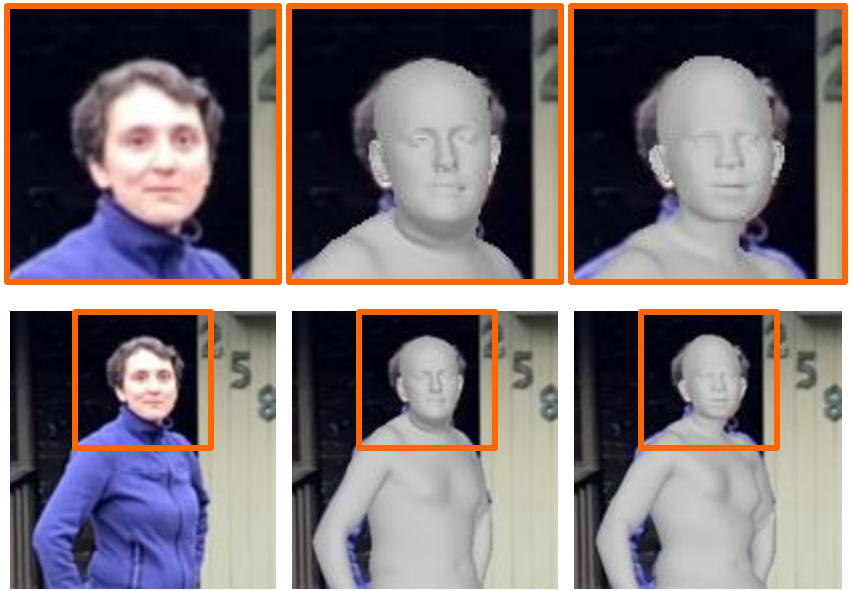
(b) Without the face loss

Mouth geometry is below the lower lip

Co-Registration of Body, Hands, and Face

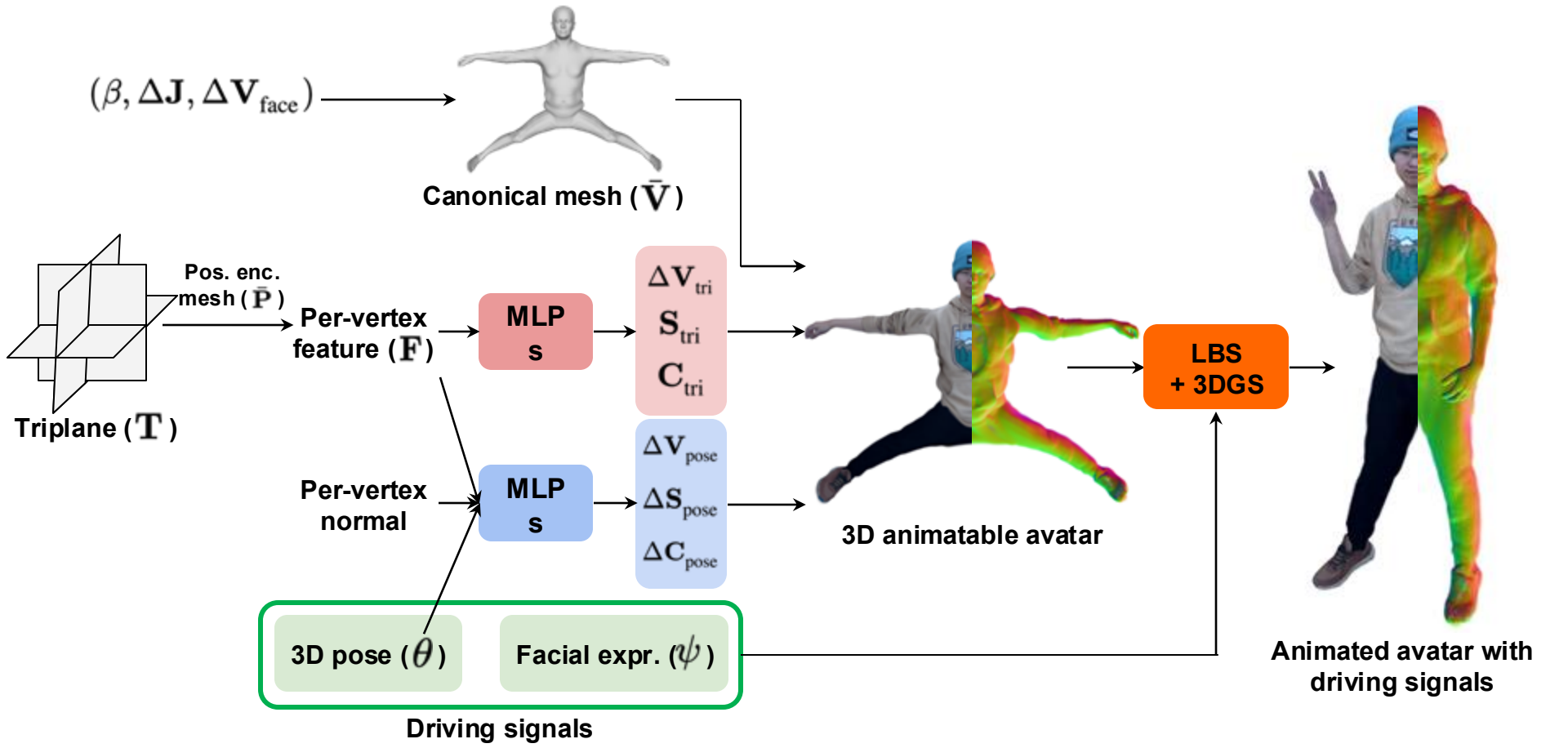


(a) Image | (b) With joint offset (Ours) | (c) Without joint offset



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

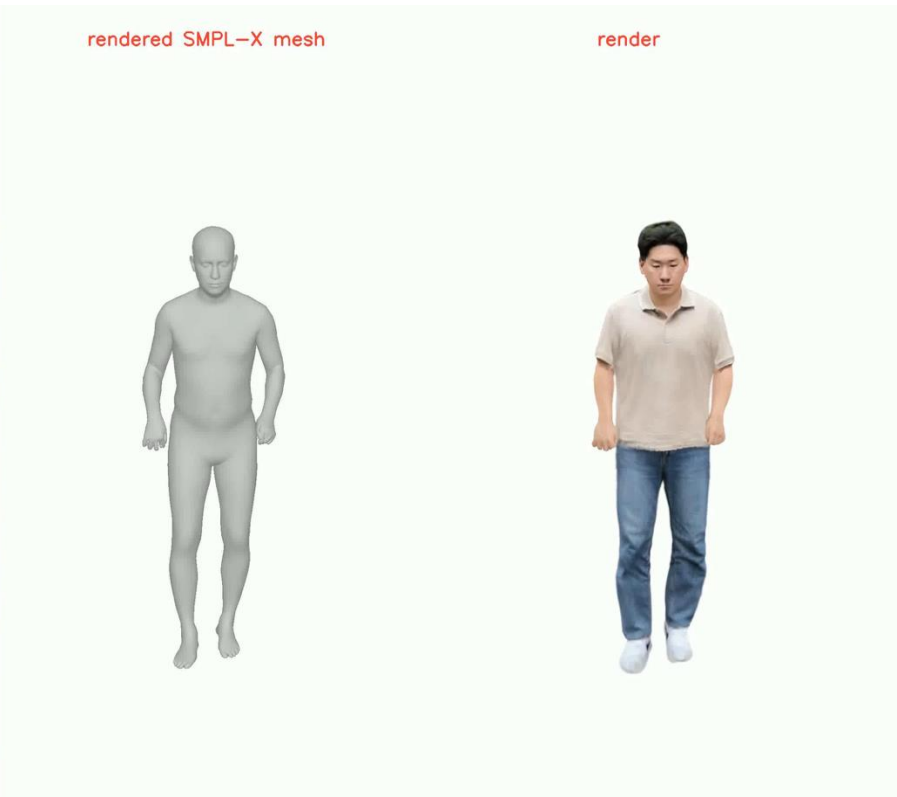
Architecture



Comparison to previous works

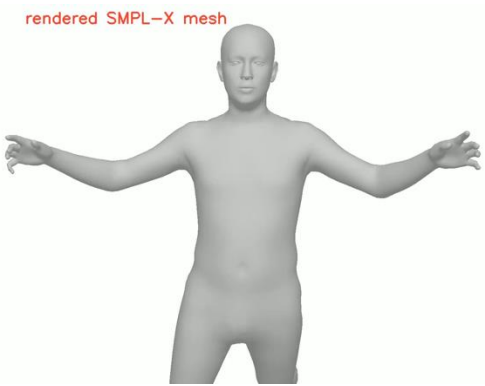
[1] Qian et al. "3DGS-Avatar: Animatable avatars via deformable 3D gaussian splatting." CVPR. 2024.

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.

Motion transfer from in-the-wild videos

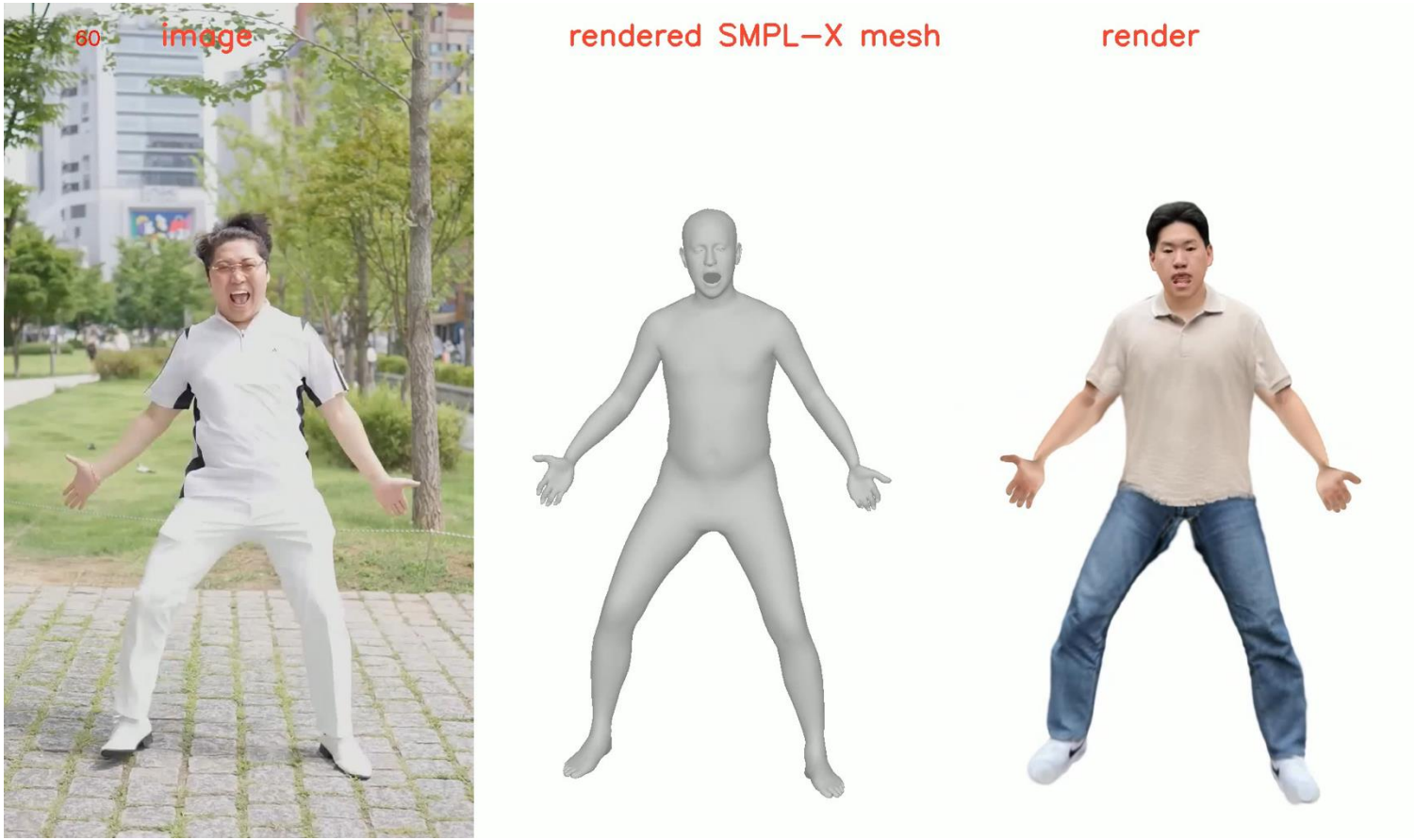


3D pose tracking is done with [1] Moon et al. "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation." CVPRW. 2022. Music is from Michael Jackson – Smooth Criminal. Dance is from YOOTAEYANG of SF9.

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



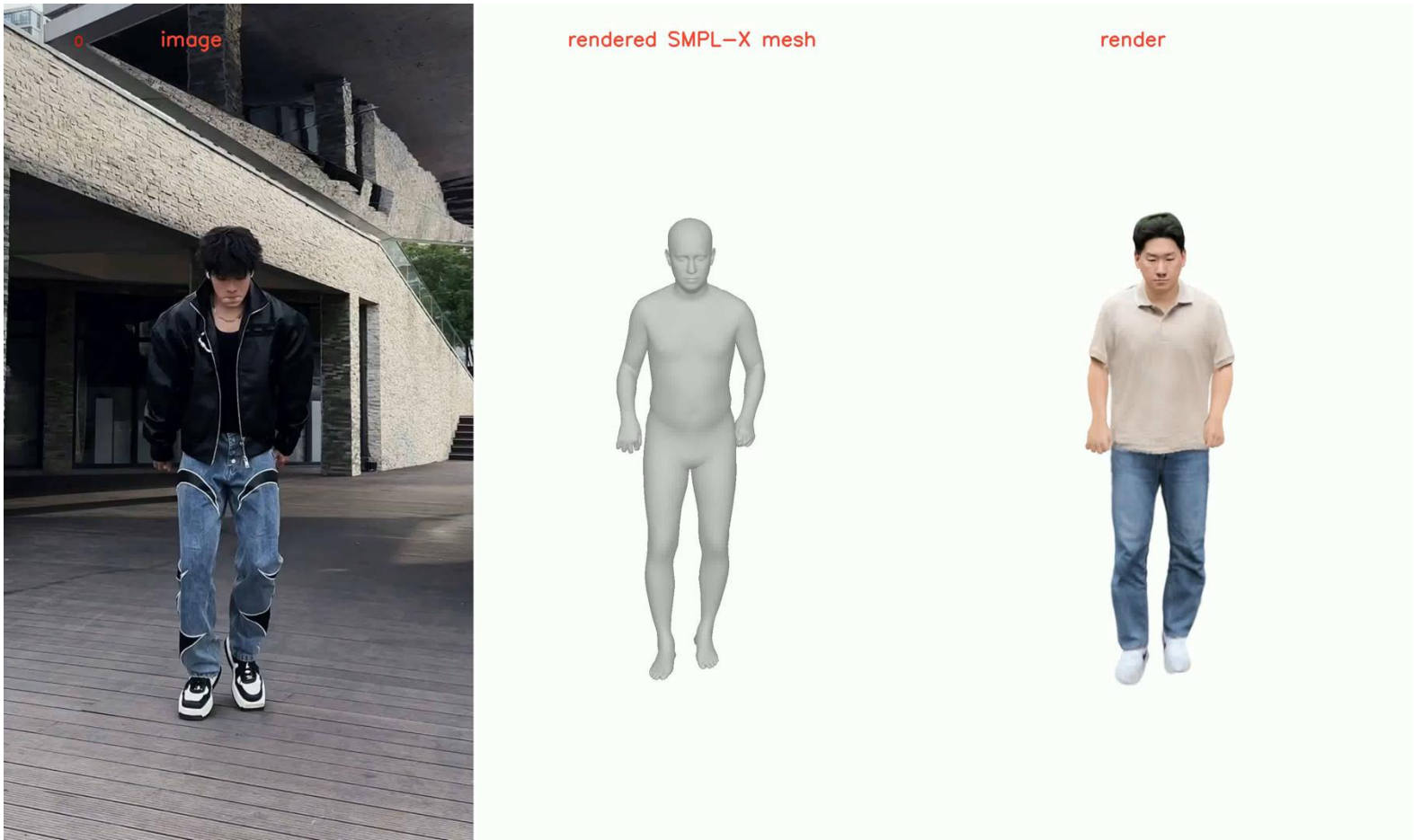
3D pose tracking is done with [1] Moon et al. "Accurate 3D hand pose estimation for whole-body 3D human mesh estimation." CVPRW. 2022. Music and video are from Youtuber the myeonsang.

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



Ph.D. DISSERTATION

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

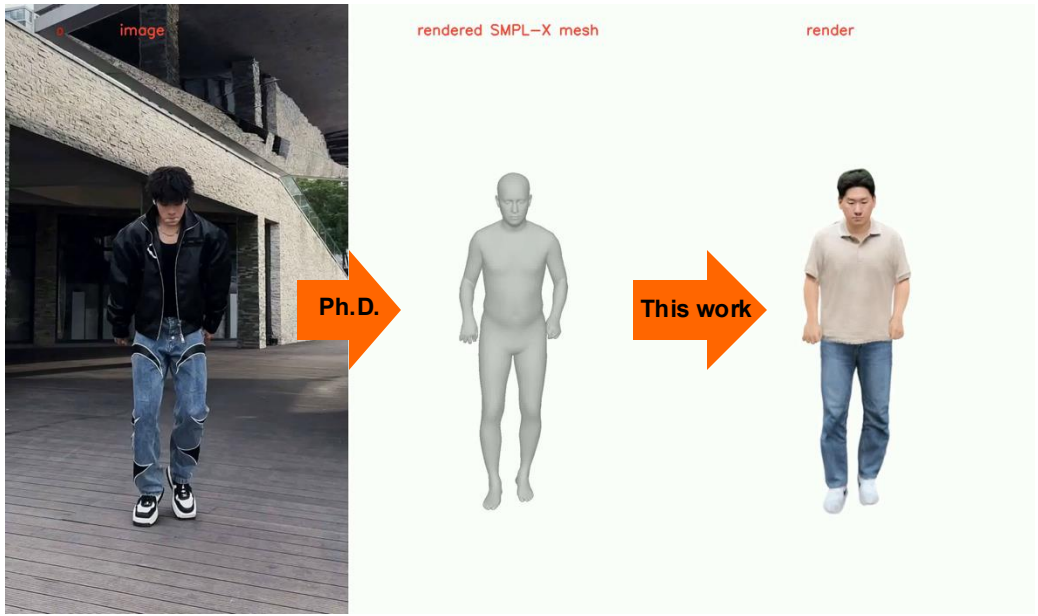
단일 이미지로부터 여러 사람의 표현적 전신 3D 자세 및 형태 추정

BY

GYEONGSIK MOON

February 2021

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



Limitations

- Actually, too many limitations ☺
- Requires a video
  - Although it’s short, still heavy requirements from users
- Limited non-rigid deformations
  - Pose-dependent cloth deformations
  - Motion-dependent cloth deformations
- Baked-in lights and shadows

Limitations

- Actually, too many limitations ☺
- Requires a video
  - Although it's short, still heavy requirements from users
- Limited non-rigid deformations
  - Pose-dependent cloth deformations
  - Motion-dependent cloth deformations
- Baked-in lights and shadows

BREAK

Any questions?

Prior matters

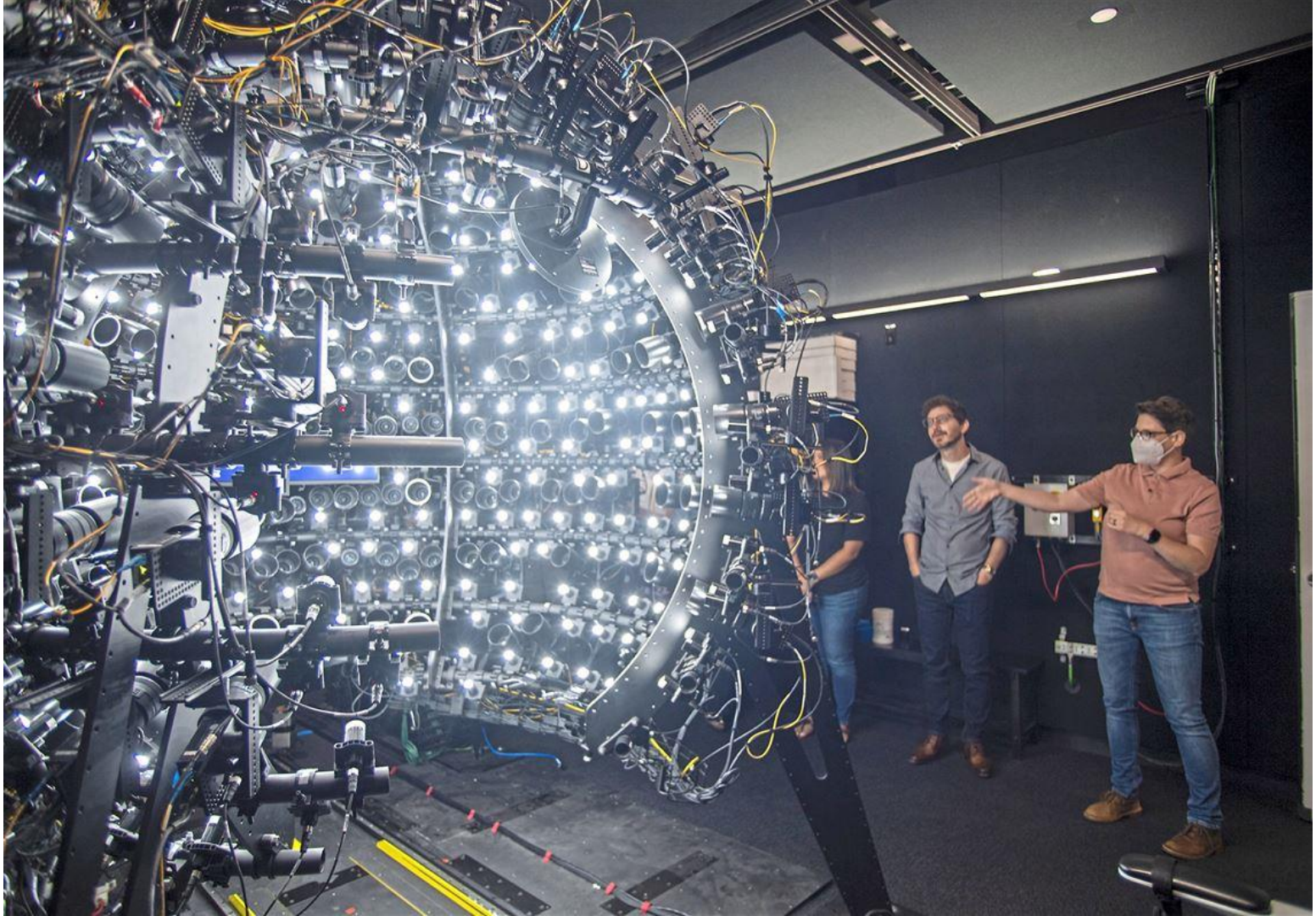
- Previous works (CV+CG-based methods) gradually changed environments from multi-view capture setup to casual ones.
- However, they still assume almost all human surfaces are visible in videos
- What if we only have a single image?
- How can we achieve viewpoint/cloth dynamics/light generalization given limited observations (e.g., a single image)?
- Prior matters!

Possible approach 1: train on large 3D datasets

- Train a large neural network on large 3D datasets (e.g., DNA-Rendering. ICCV 2023)
- Widely used direction for previous 10 years
    - Train a network on large 3D datasets and test on their testing splits
- It’s quite expensive to capture all possible combinations of 3D data (# of subjects, # of poses, # of lights, # of viewpoints, …)
    - It makes diversity of such 3D datasets limited…

Possible approach 1: train on large 3D datasets

- IDOL (CVPR 2025), AniGS (CVPR 2025), LHM (ICCV 2025), ...
- Train a large feedforward network on large 3D datasets
- They are good, but cannot represent all human physical attributes
    - Viewpoints, poses, cloth dynamics, lights, …
    - Capturing 3D datasets with diverse (viewpoints, poses, clothes, lights) makes it very expensive and not scalable
    - Reality Labs Research at Meta is (was) pushing this direction, but not sure…



Possible approach 2: pre-trained generative models

- Image/video/multi-modal generative models are trained on any images/videos/audio data without requiring full 3D data (i.e., no 3D geo/tex/cam/lights)
- Hence, such generative models can be trained on highly diverse and huge data
- Great prior modeling capability thanks to highly diverse and huge training data
- It makes something!

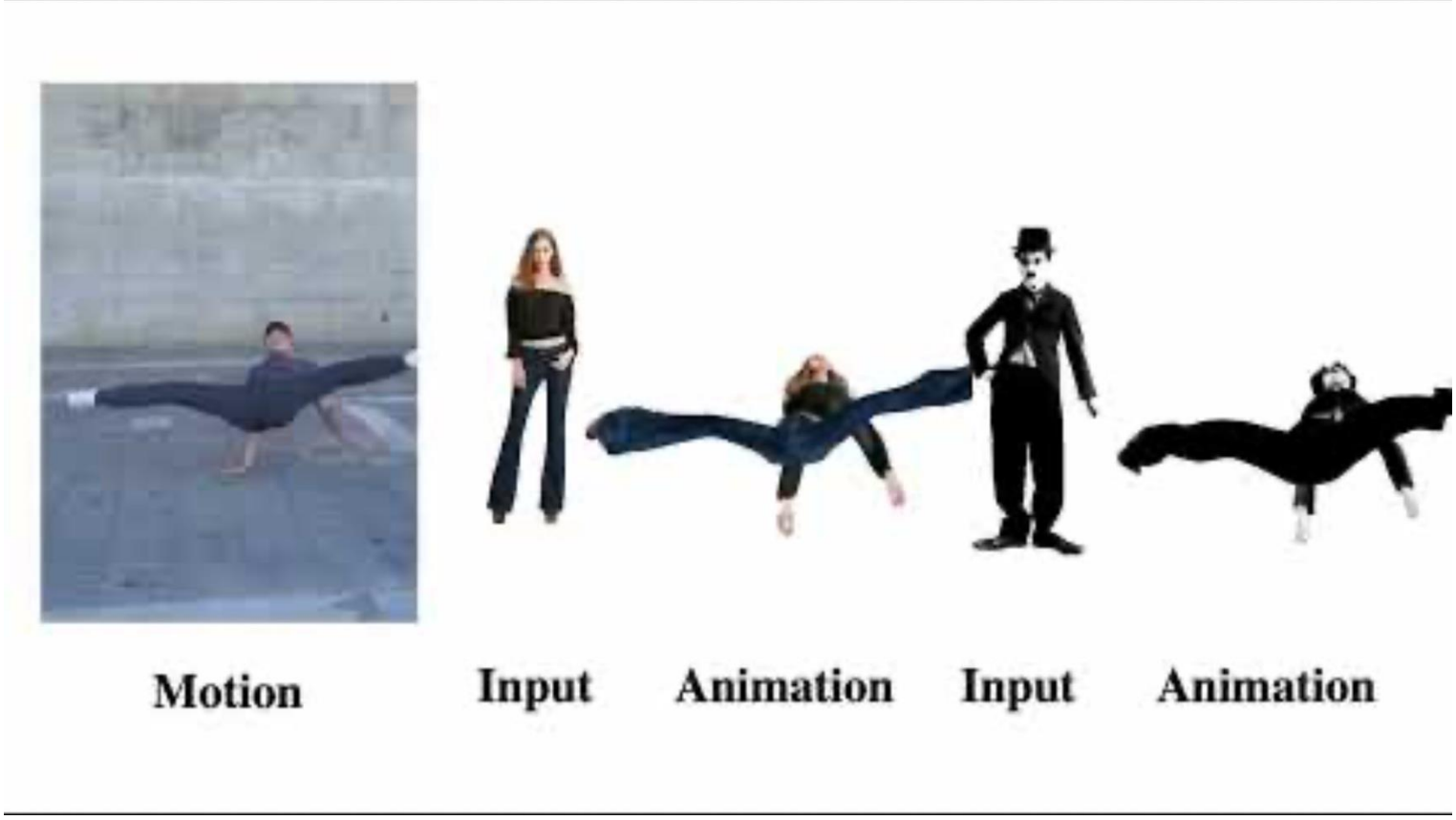
Possible approach 2: pre-trained generative models

- There are generative models for human animations
- Mostly, ControlNet-style
- Takes a single image and controlling signal (target 2D pose sequences for the animation)
- Just generative animated videos without any 3D geo/tex/cam/lights
- Very recent work from ByteDance: https://byteaigc.github.io/X-Unimotion/

What’s next?

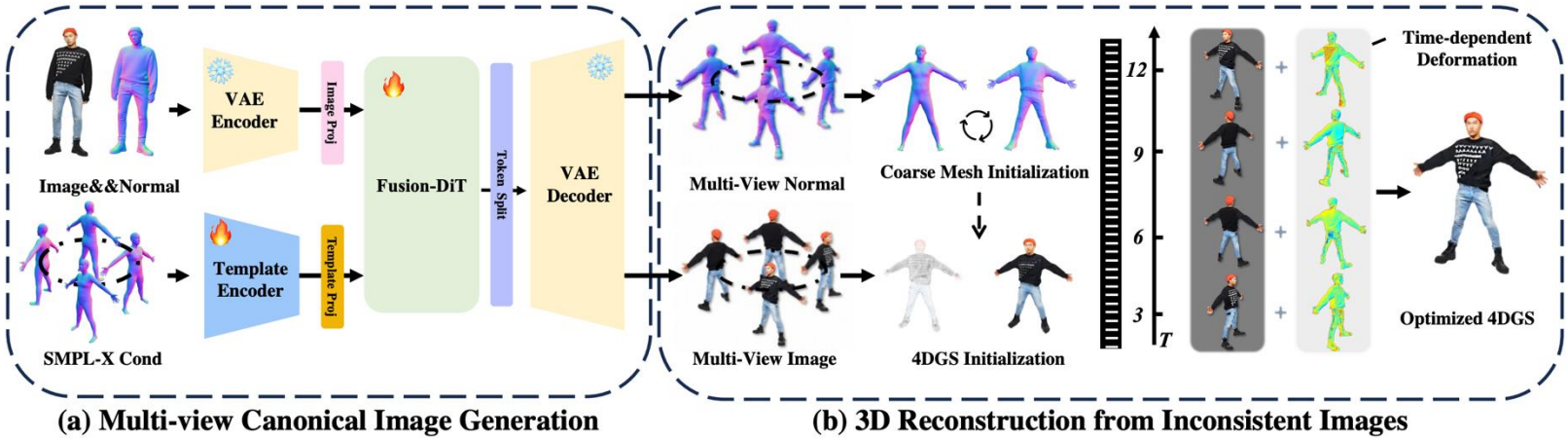
- 3D-based methods
  - Good: Consistent, easy to edit, physically plausible
  - Weak: Data collection is not scalable -> weak prior modeling capability
- Generative-based methods
  - Good: Data collection is scalable (?) -> strong prior modeling capability
  - Weak: inconsistent, hard to edit, physically implausible
- Combining them together

AniGS (CVPR 2025)



AniGS (CVPR 2025)

- Generate multi-view training images with canonical pose
- Fit 3D avatar to the generated images
- Animate with 3D poses
- Viewpoint prior is from generative model, animation prior is from SMPL-X, no light prior



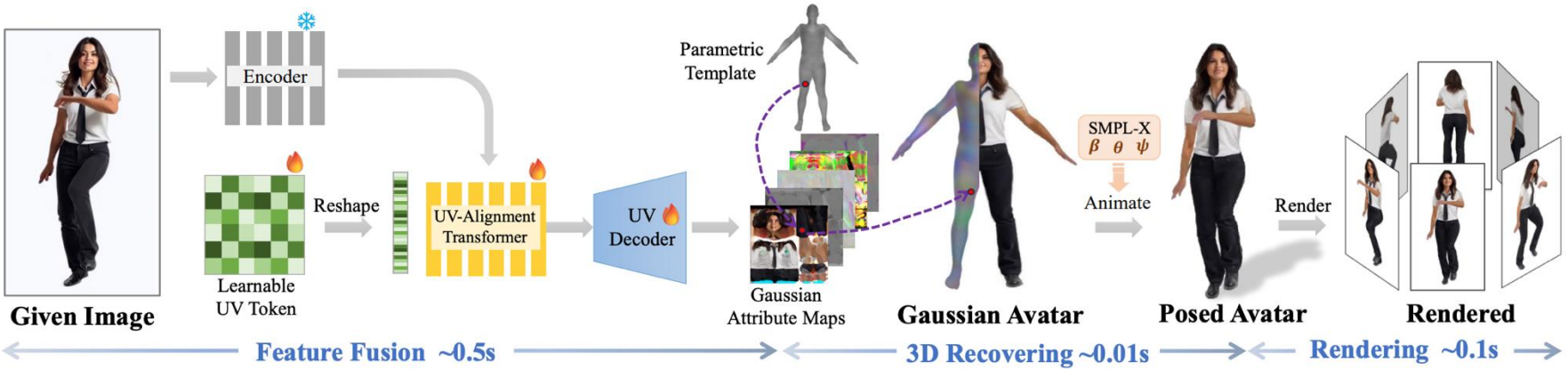
AniGS (CVPR 2025)

- Animation prior is from SMPL-X
  - No idea how clothes deform
  - Without any non-rigid deformations

Input image AniGS

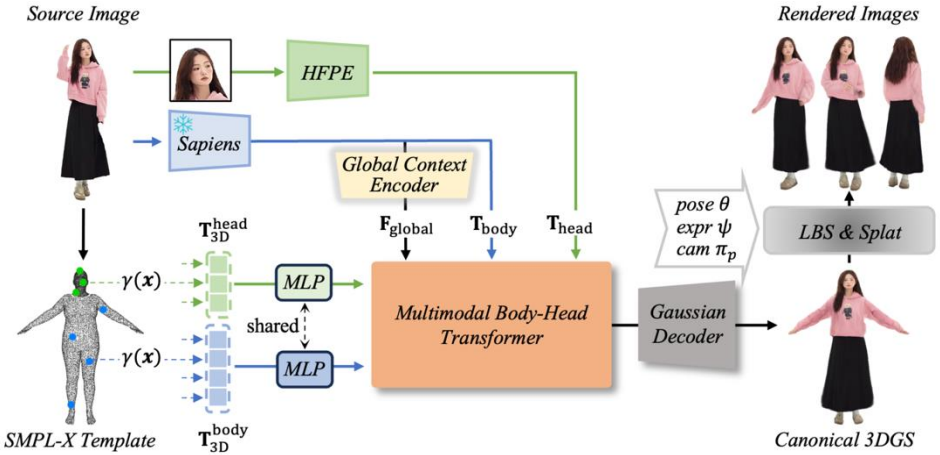
Similar 3D-based methods

- IDOL (CVPR 2025)
  - Train a large network on large synthetic 3D datasets
  - Animate with SMPL-X poses (no idea of clothing deformations)
  - Limited training data (no motion) / limited representation (no dynamics representation)



Similar 3D-based methods

- LHM (ICCV 2025)
  - Train a large network on large 3D datasets
  - Animate with SMPL-X poses (no idea of clothing deformations)
  - Limited training data (no motion) / limited representation (no dynamics representation)

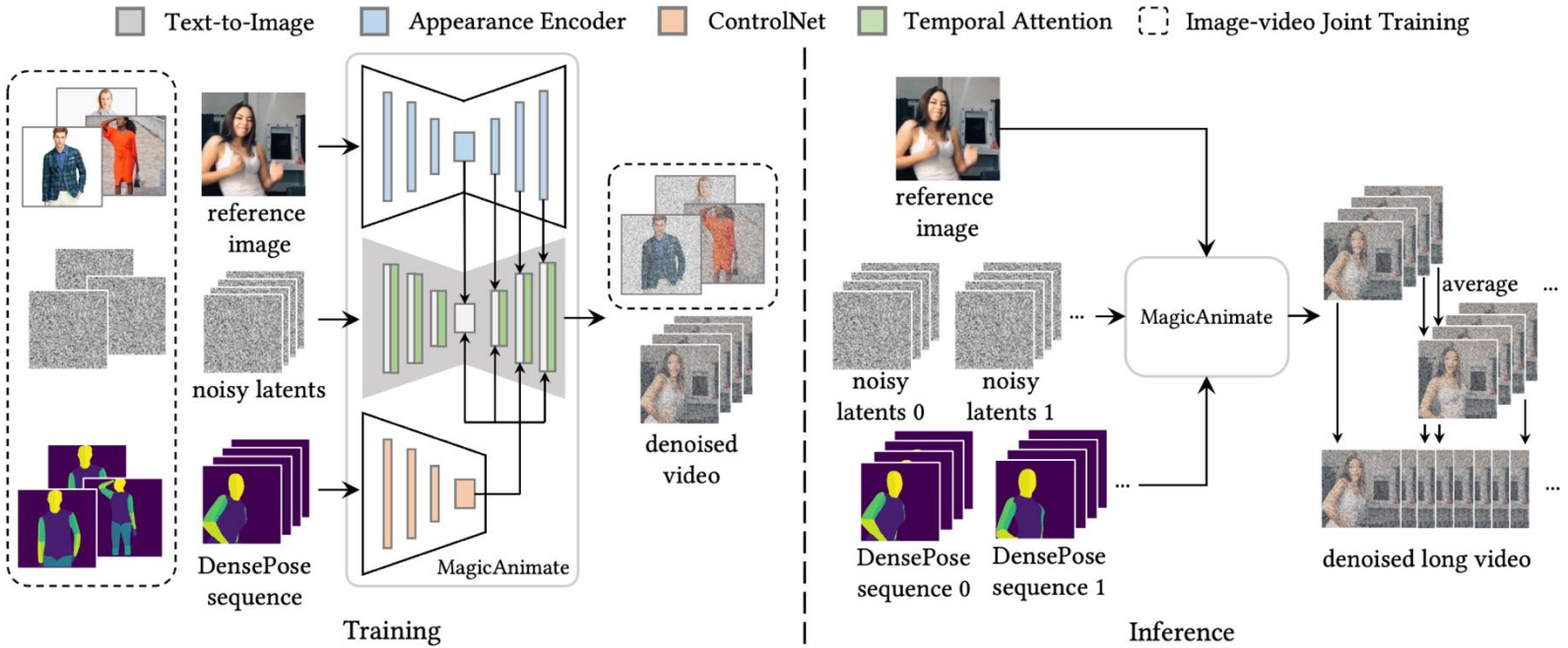


3D-based methods

- If we have enough training data, 3D-based methods should be good
    - See below Meta's results
    - In reality, we do not bring individual to capture studios ☹
- Without training data, we rely on weak priors (animation prior of SMPL-X), which makes the quality not good…

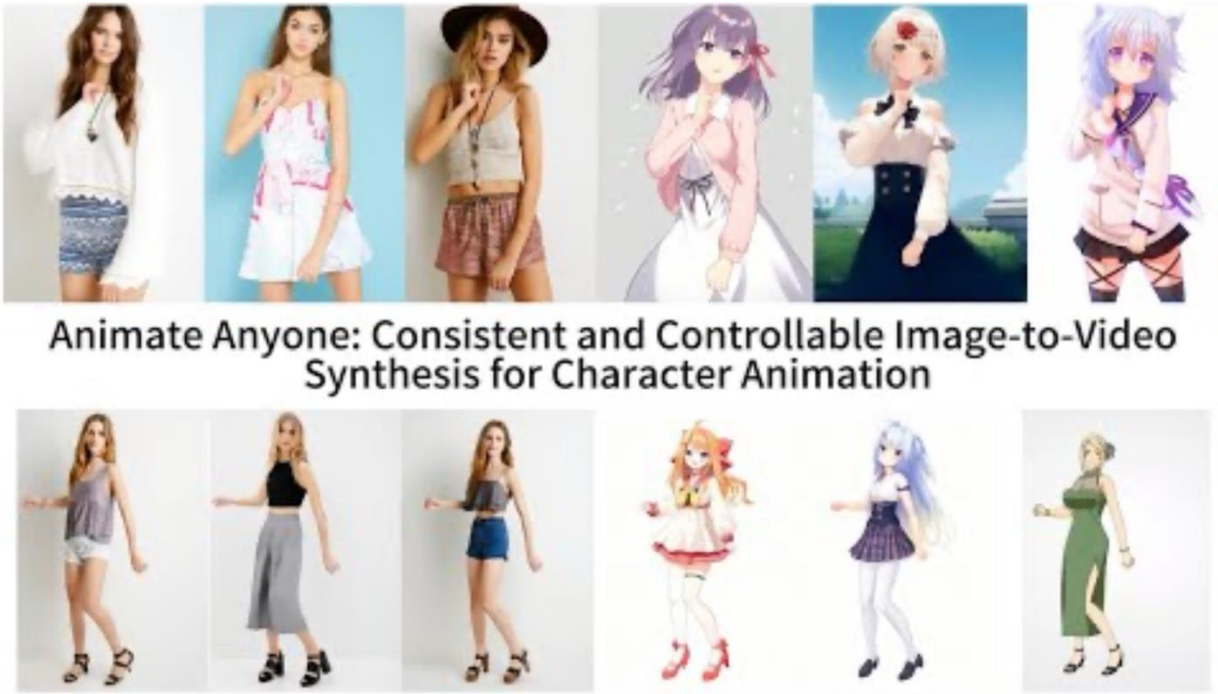
Generative methods

- MagicAnimate (CVPR 2024. https://showlab.github.io/magicanimate/)



Generative methods

- AnimateAnyone (CVPR 2024)



Generative methods

- AnimateAnyone (CVPR 2024)
- https://drive.google.com/file/d/1F79coHx61l0vrVZLcTrQAKYM0CpXnDtO/view?usp=sharing
- Given this quality, I thought generative methods are pretty far behind the 3D-basd methods

Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation

Li Hu
Institute for Intelligent Computing, Alibaba Group
hooks.hl@alibaba-inc.com

Single author in CVF version?

https://humanaigc.github.io/animate-anyone/

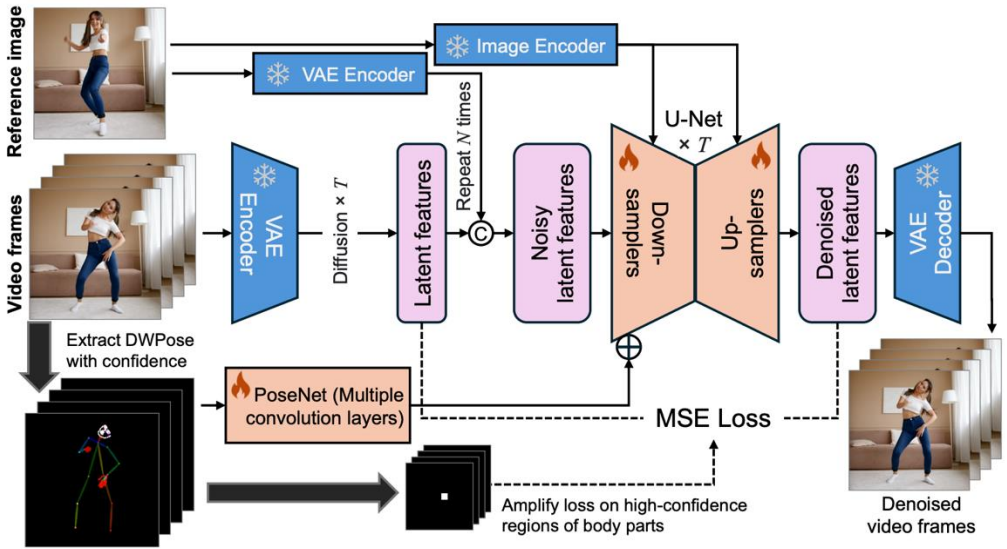
Animate Anyone: Consistent and Controllable Image-to-Video Synthesis for Character Animation

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, Liefeng Bo
Institute for Intelligent Computing, Alibaba Group
{hooks.hl, zimu.gx, futian.zp, xisheng.sk, zhangbang.zb, liefeng.bo}@alibaba-inc.com

https://humanaigc.github.io/animate-anyone/

MimicMotion (ICML 2025)

- Modulate pre-trained video generative models with 2D motion features
- Now some reasonable quality (https://tencent.github.io/MimicMotion/)



MimicMotion (ICML 2025)

UniAnimate-DiT

- https://github.com/ali-vilab/UniAnimate-DiT
- Mostly similar to previous generative methods
    - Modulate pre-trained video generative models with driving signal (2D motion)
    - Fine-tune with more data
    - Better pre-trained video generative models

ICCV OCT 19-23, 2025 HONOLULU HAWAII

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

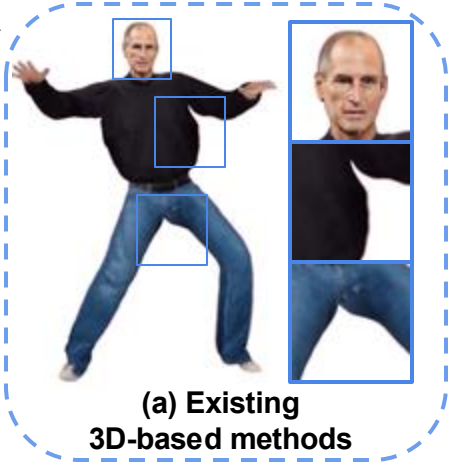
Geonhee Sim

Gyeongsik Moon

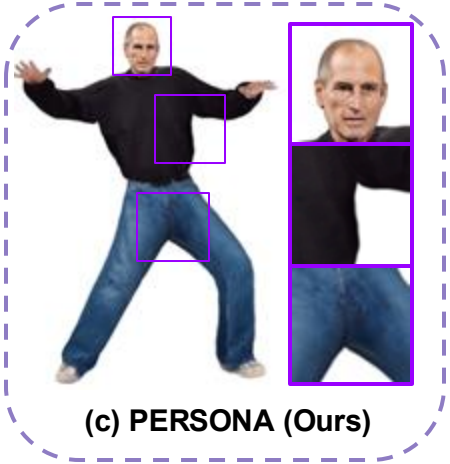
KOREA UNIVERSITY 1905 Visual Computing and AI Lab

Input image

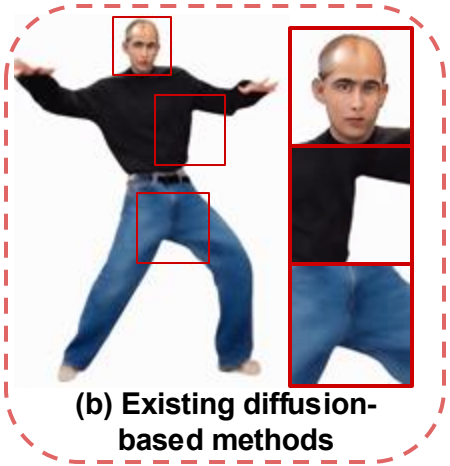
Authenticity



(a) Existing
3D-based methods



(c) PERSONA (Ours)



(b) Existing diffusion-
based methods

Pose-driven deformations

What’s next? (slide from previous page)

- 3D-based methods
  - Good: Consistent, easy to edit, physically plausible
  - Weak: Data collection is not scalable -> weak prior modeling capability
- Generative-based methods
  - Good: Data collection is scalable (?) -> strong prior modeling capability
  - Weak: inconsistent, hard to edit, physically implausible
- Combining them together

Training data for PERSONA

Input image

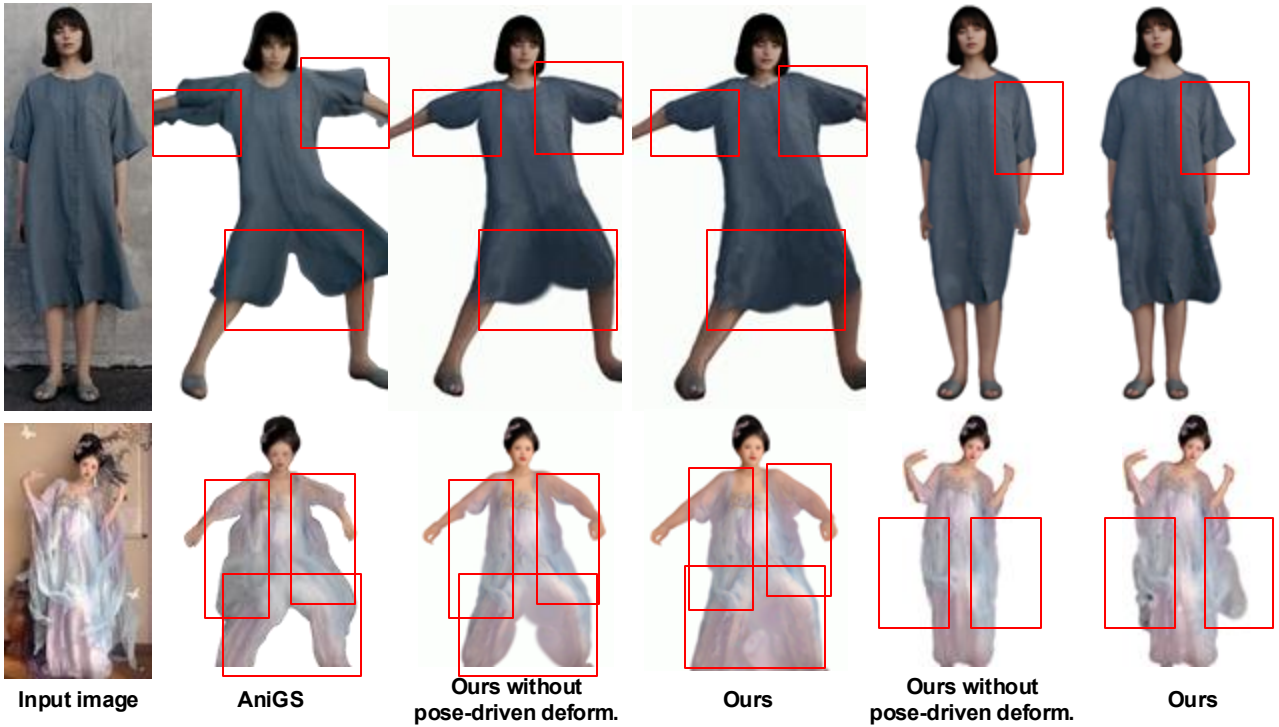
Diffusion-based animator

Generated pose-rich videos

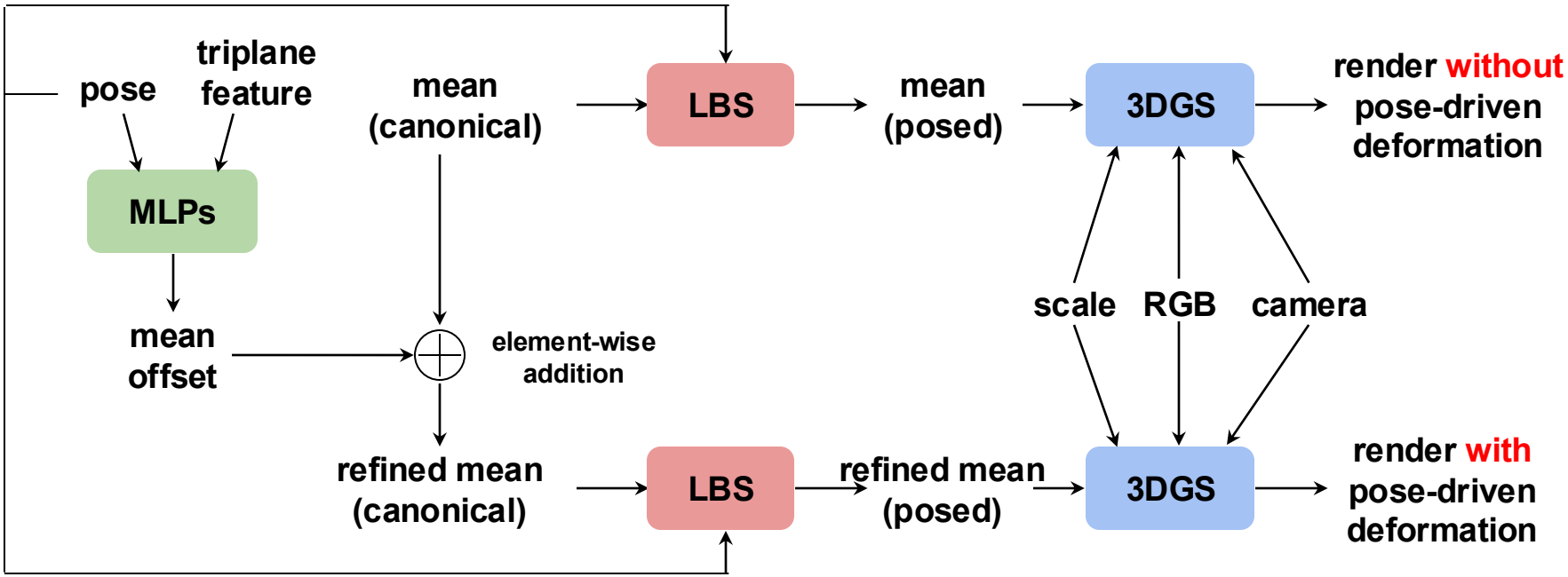
PERSONA

- Obtain training data with diverse motion from generative methods
- Includes how clothes deform with diverse motions
- Previous 3D-based methods have no idea how clothes deform with diverse motions
- Now, we can represent pose-driven cloth deformations!

Non-rigid pose-driven cloth deformations



Non-rigid pose-driven cloth deformations

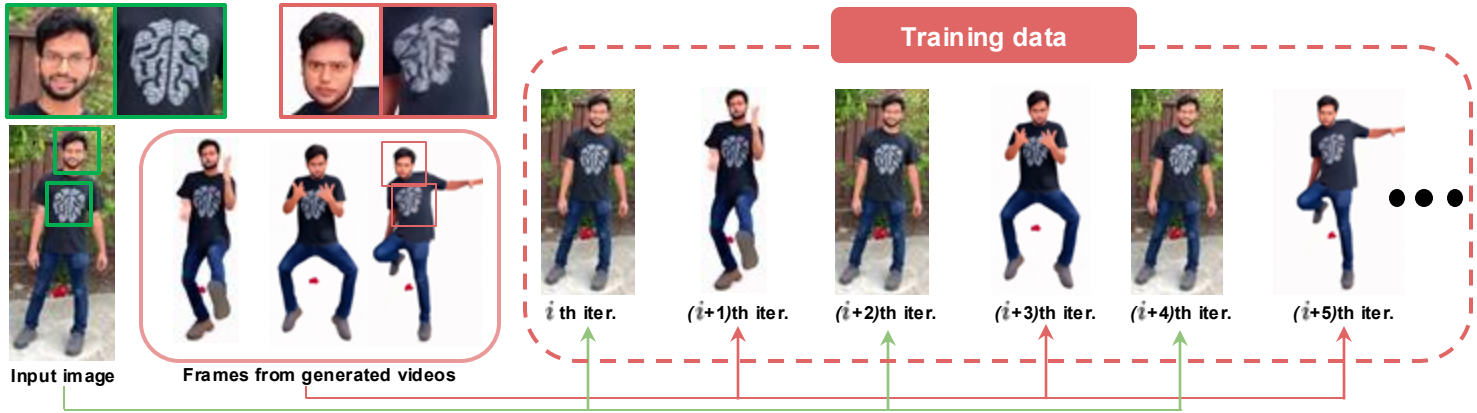


PERSONA

- Obtain training data with diverse motion from generative methods
- Includes how clothes deform with diverse motions
- Previous 3D-based methods have no idea how clothes deform with diverse motions
- Challenge: inconsistent and noisy generative videos…
    - If such generated videos are consistent and clean (real videos), we have enough observations and can use any previous 3D-based works
    - However, generative models are inherently inconsistent and noisy although they're rapidly getting better

Balanced sampling

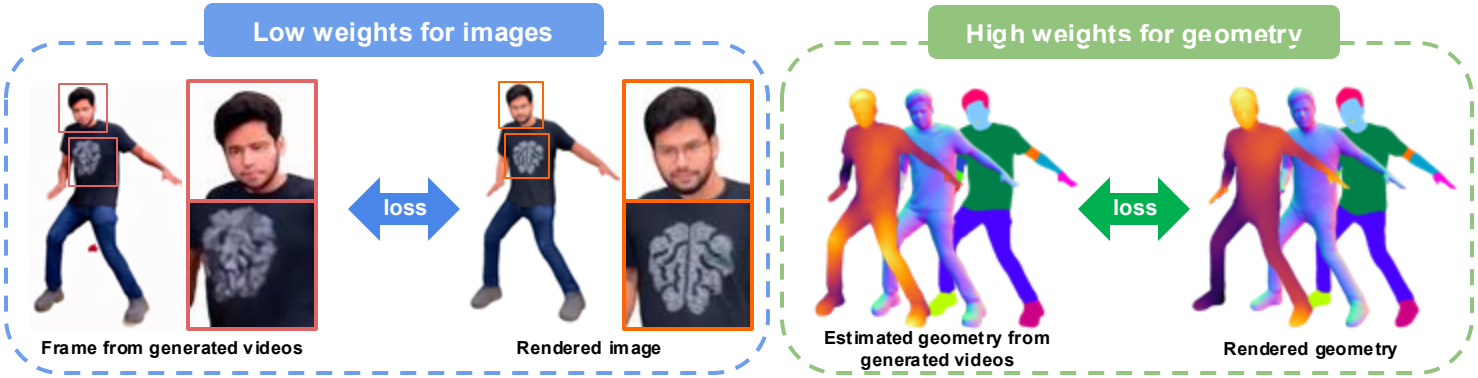
- Authenticity: preserving identity of the person in the input image
- Necessary as we're aiming for personalized representation
- Just oversample the input image during the optimization



(a) Balanced sampling to preserve identity for personalization

Geometry-weighted optimization

- Generated videos have inconsistent and noisy textures
- Simply optimizing 3D avatar on such noisy texture results in blurry texture
- We use low weights for image loss and high weights for geometry loss
- Geometry is relatively robust to such noise in generated videos!



(b) Geometry-weighted optimization for sharp renderings in pose-driven deformations

Balanced sampling

- Balanced sampling is necessary to keep the identity (high authenticity) and sharp texture of the input image
- Additional albedo loss is effective to remove baked-in shadow artifacts

(a) Input image

(b) Ours wo. balanced sampling

(c) Ours wo. albedo loss

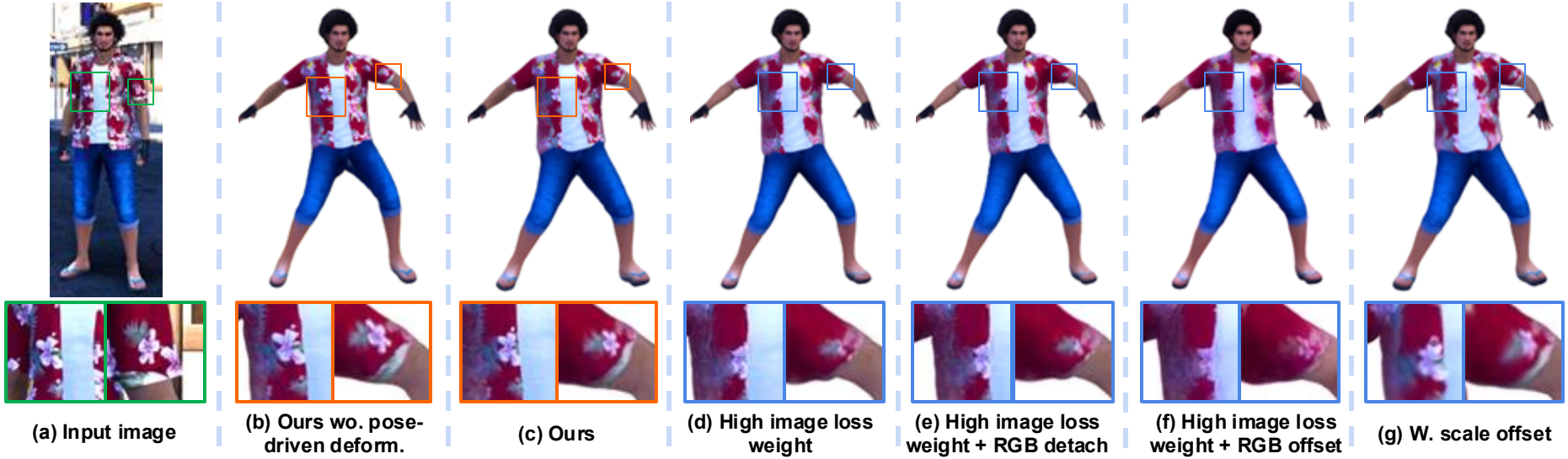
(d) Ours

(e) 4DGS

(f) 4DGS + our balanced sampling

Geometry-weighted optimization

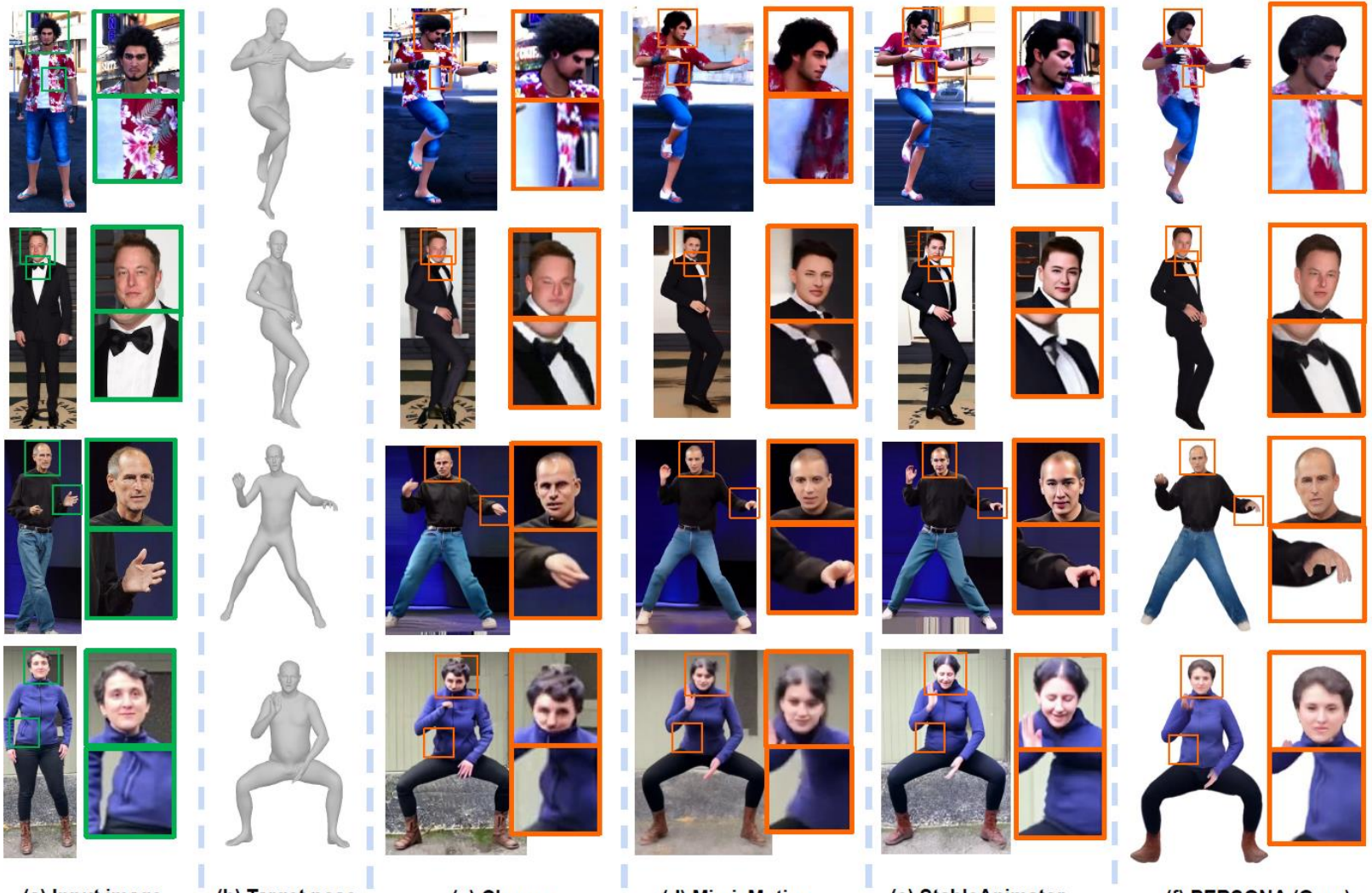
- Geometry-weighted optimization is necessary to achieve sharp textures while modeling non-rigid cloth deformations



(a) Input image | (b) Ours wo. pose-driven deform. | (c) Ours | (d) High image loss weight | (e) High image loss weight + RGB detach | (f) High image loss weight + RGB offset | (g) W. scale offset


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)

Limitations

- Motion-dependent cloth dynamics (WIP)
    - We only modeled pose-dependent cloth dynamics
- Separated and multi-layer cloth representation (WIP)
    - Currently, all Gaussians starts from SMPL-X template
- No light modeling
- Relatively lower quality than SOTA video generative models (Veo3)
    - Our training data is from generative methods, but most open-sourced generative models do not have good quality ☹

Cloth dynamics

- PGC (CVPR 2025. physics simulation-based representation)
  - https://phys-gaussian-cloth.github.io/
  - Simulations are prone to noisy 3D poses during the animation…
  - Not sure about real-time animation capability
  - Require multi-view capture setup with a simple pose

What’s next?

- Video generative models are really cool
- They have quite strong prior modeling capability
- From my graduate school, I’ve read papers for prior modeling using 3D data, but their capability is still limited
- Video generative models can use **any** forms of data without requiring expensive 3D data capture
- They’re rapidly getting better

What’s next? (slide from previous page)

- 3D-based methods
  - Good: Consistent, easy to edit, physically plausible
  - Weak: Data collection is not scalable -> weak prior modeling capability
- Generative-based methods
  - Good: Data collection is scalable (?) -> strong prior modeling capability
  - Weak: inconsistent, hard to edit, physically implausible
- Combining them together

What’s next?

- But some day, maybe 3D representation could become not necessary..?
- What if we can do everything in a latent space of generative models?
- At this moment, 3D representation seems still useful, but not sure after 5 or 10 years

Our group

- During graduate school, 3D human pose estimation from a single image was my main research area
- After getting Ph.D., I’m gradually changing my area from 3D human pose estimation to more diverse area
    - Human representation
    - Human video generative models
- For the 3D human pose estimation, I’m mostly focusing on physically plausibility for robotics
    - Regression says nothing about physical plausibility

- Fresh lab (started this March)
- We're getting bigger
- https://www.vcai.korea.ac.kr
- Welcome any collaborations!

Master Students

Geonhee Sim
kh6362@dgist.ac.kr

Undergraduate Interns

Minseo Kang
danielk0112@korea.ac.kr

Changyeop Lee
ckd248@korea.ac.kr

Joohyun Kwon
rnjswngus00@dgist.ac.kr

Gaeun Ko
jjsk3218@korea.ac.kr

Taejun Roh
rohtj_312@dgu.ac.kr

Taekwan Kim
xorhks0622@korea.ac.kr

Myungkyung Shin
2022240100@korea.ac.kr

Hyobeom Kim
i1t28ly@dgu.ac.kr


Understand humans, and
eventually, interact with humans