Human 3D Reconstruction and Generation

Zhuo Su

总览

Overview

1. 背景

1. Background

1.1 Problem Definition

1.2 VR/AR Applications

2. 传统方法到神经渲染

2. From Volumetric Capture to Neural Rendering

2.1 Human Volumetric Capture

2.2 Human Neural Rendering

3. 先验模型到三维生成

3. From Prior Model to 3D Generation

3.1 Animatable Avatar Creation

3.2 Human 3D Generation

1. 背景

1. Background

1.1 问题定义

1.1 Problem Definition

1.2 应用背景

1.2 VR/AR Applications

1.1 问题定义

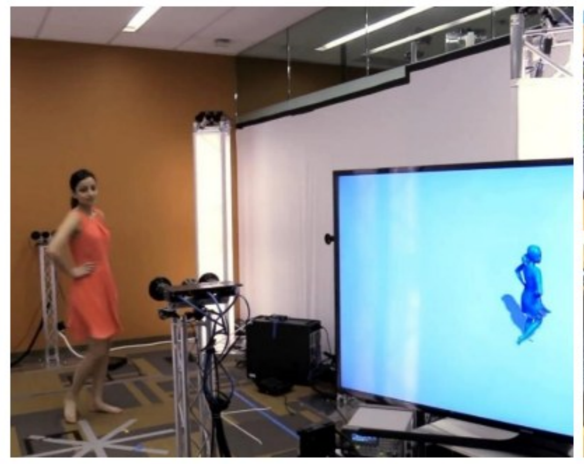
1.1 Problem Definition

Human Reconstruction and Generation | How to accurately reconstruct or generate 3D human models from image or video inputs , including static reconstruction/generation (3D), dynamic reconstruction (4D), and animatable avatar creation.

USC PIFu

Bytedance HumanSplat

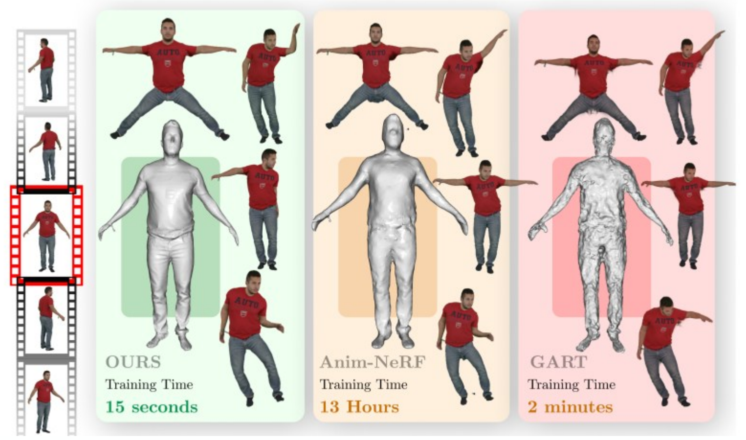
UW DynamicFusion



Microsoft Fusion4D

Google Motion2Fusion

TUM GaussianAvatars



TU iHuman

Static Reconstruction/Generation

Dynamic Reconstruction

Animatable Avatar Reconstruction

ByteDance字节跳动

1. 背景

1. Background

ByteDance字节跳动

1.2 VR/AR应用背景

1.2 VR/AR Applications

Immersive Communication｜三维通信

Connecting people through lifelike 3D reconstruction for enhanced interaction

Volumetric Video｜体积视频

Delivering realistic human performance with dynamic capture, live streaming, and playback

3D Portrait/Gaming/Fashion Design ｜三维画像/游戏/服装设计

Expanding creativity and immersion through diverse 3D human modeling applications in VR/AR

ByteDance字节跳动

总览

Overview

1. 背景

1. Background

1.1 Problem Definition

1.2 VR/AR Applications

2. 传统方法到神经渲染

2. From Volumetric Capture to Neural Rendering

2.1 Human Volumetric Capture

2.2 Human Neural Rendering

3. 先验模型到三维生成

3. From Prior Model to 3D Generation

3.1 Animatable Avatar Creation

3.2 Human 3D Generation

传统方法	神经渲染	先验模型	三维生成	?	?
2018.07 – 2021.06	2021.07 – now	2023.05 – now	2023.06 – now		
4D Volumetric Capture	4D Neural Rendering	Avatar creation via Prior model	Human 3D Generation		

The Past

The Present

The Future

ByteDance字节跳动

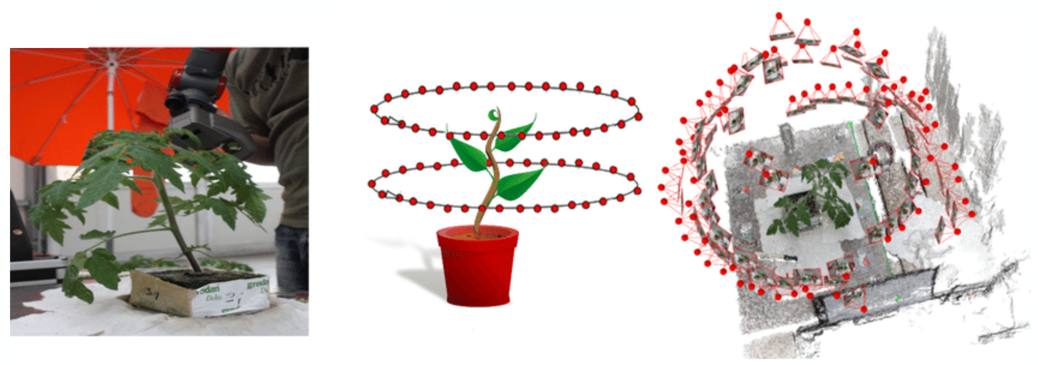
2. 动态重建：从传统方案到神经渲染

2. Dynamic Reconstruction：From Volumetric Capture to Neural Rendering

2.1 传统方案：体积捕获

2.1 Human Volumetric Capture

4D Reconstruction | 3D coordinate + 1D Temporal Motion



Geometry Reconstruction

3D

Static Scene

Canonical Geometry

3D

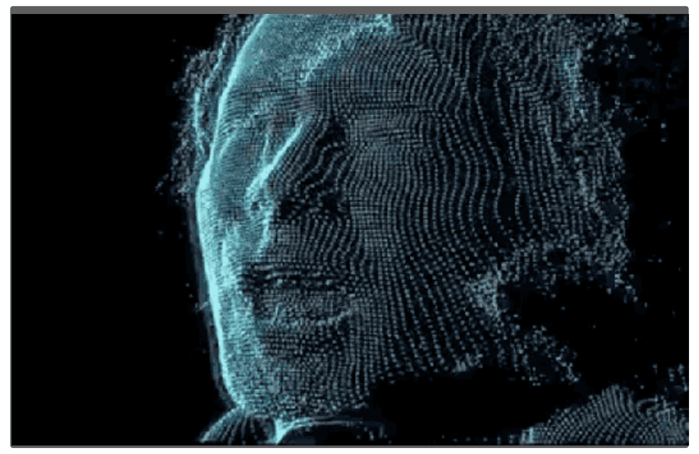
+

Temporal Motion

1D

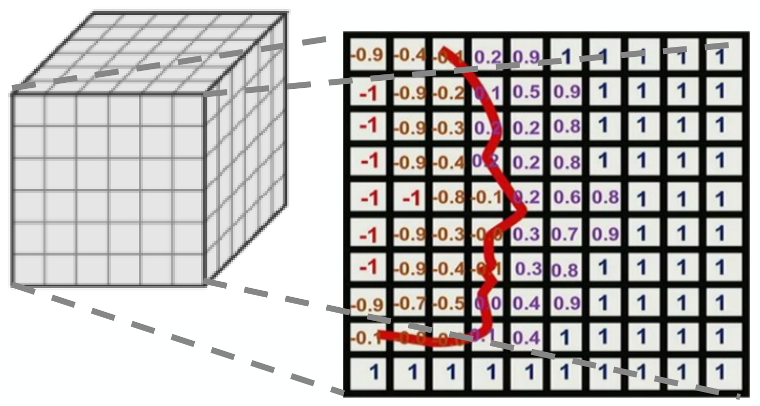
Dynamic Scene

Volumetric Capture | Canonical Model + Motion Field



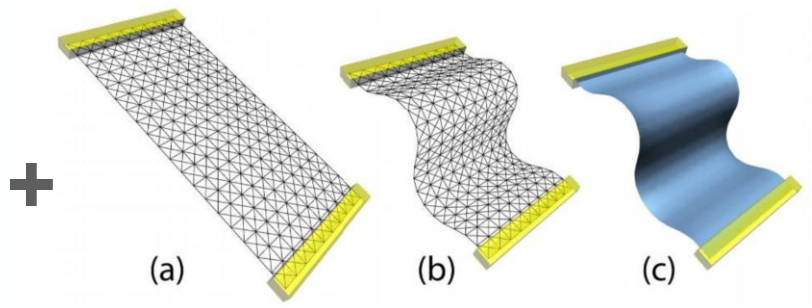
Volumetric Capture

=



TSDF Fusion

+



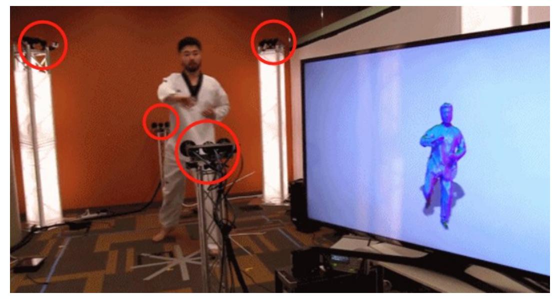
Embedded Deformation Graph

2.1 传统方案：体积捕获

2.1 Human Volumetric Capture

Challenges#1

Multiple cameras rely on fixed spatial positions, making calibration and synchronization cumbersome.



8 custom RGBD cameras consisting of 24 individual cameras.

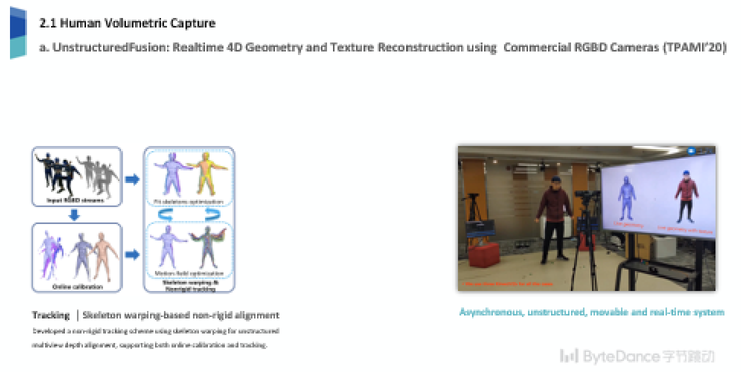
Challenges#2

The performer needs to rotate fully, and it is not robust for fast or complex motion tracking.

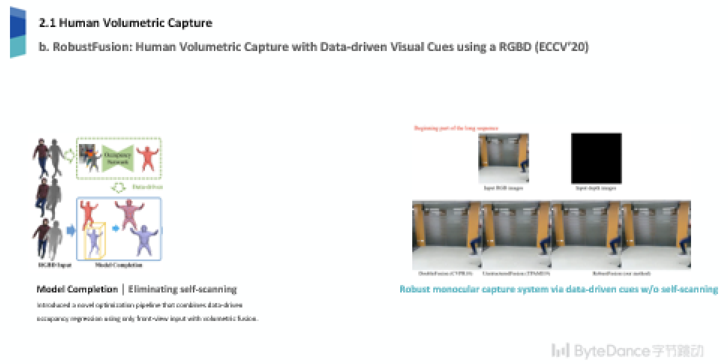
Challenges#3

It is challenging to robustly reconstruct human-object interaction scenes with complex motions.

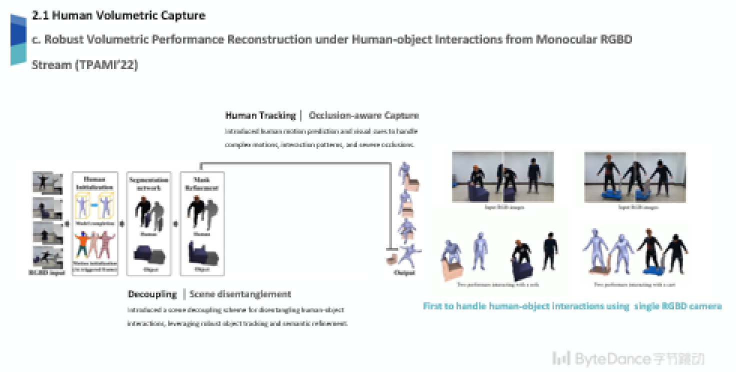
Structured & inflexible



Self-scanning at single-view

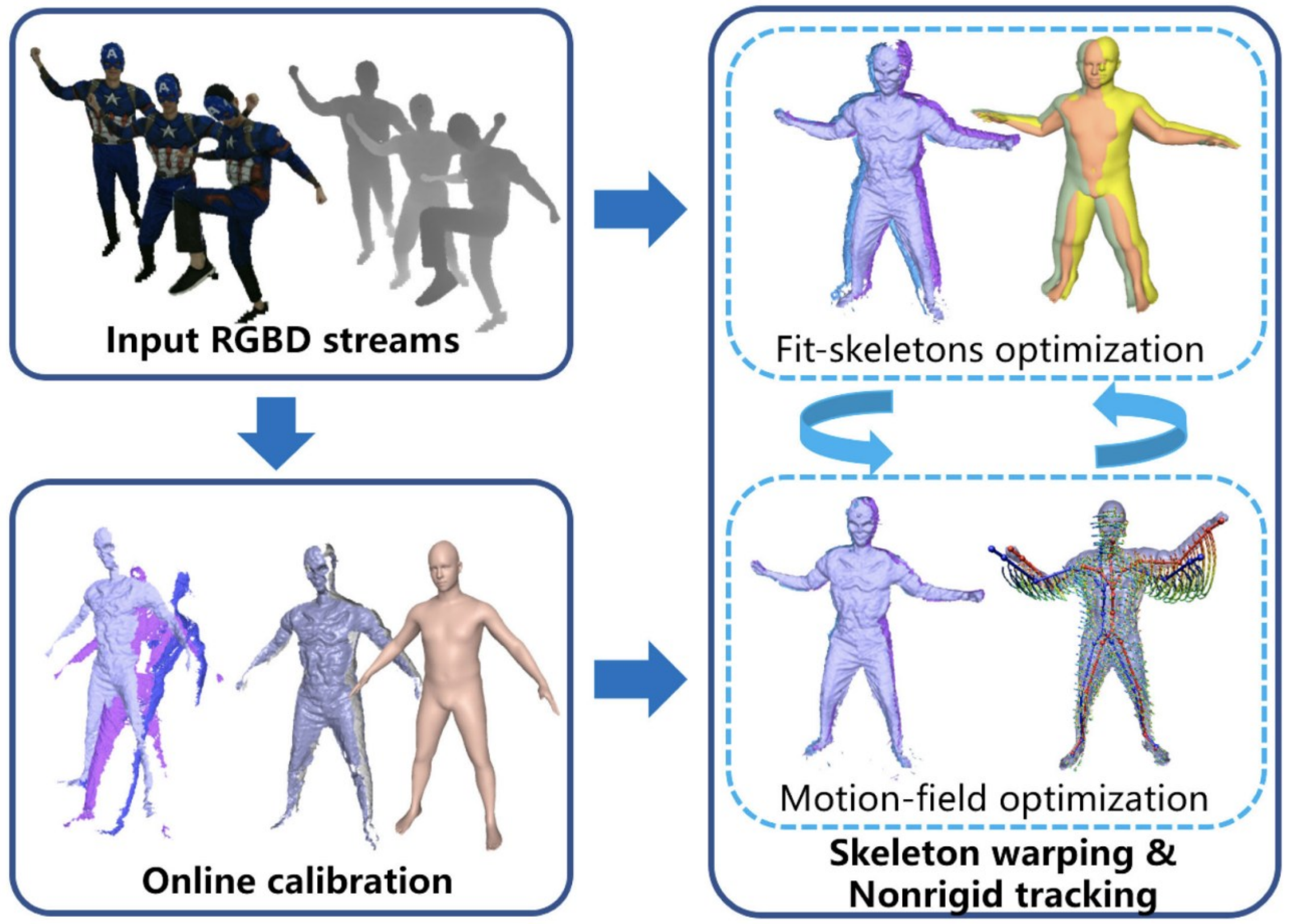


Human-object interaction



2.1 Human Volumetric Capture

a. UnstructuredFusion: Realtime 4D Geometry and Texture Reconstruction using Commercial RGBD Cameras (TPAMI’20)



Tracking | **Skeleton warping-based non-rigid alignment**

Developed a non-rigid tracking scheme using skeleton warping for unstructured multiview depth alignment, supporting both online calibration and tracking.

Asynchronous, unstructured, movable and real-time system

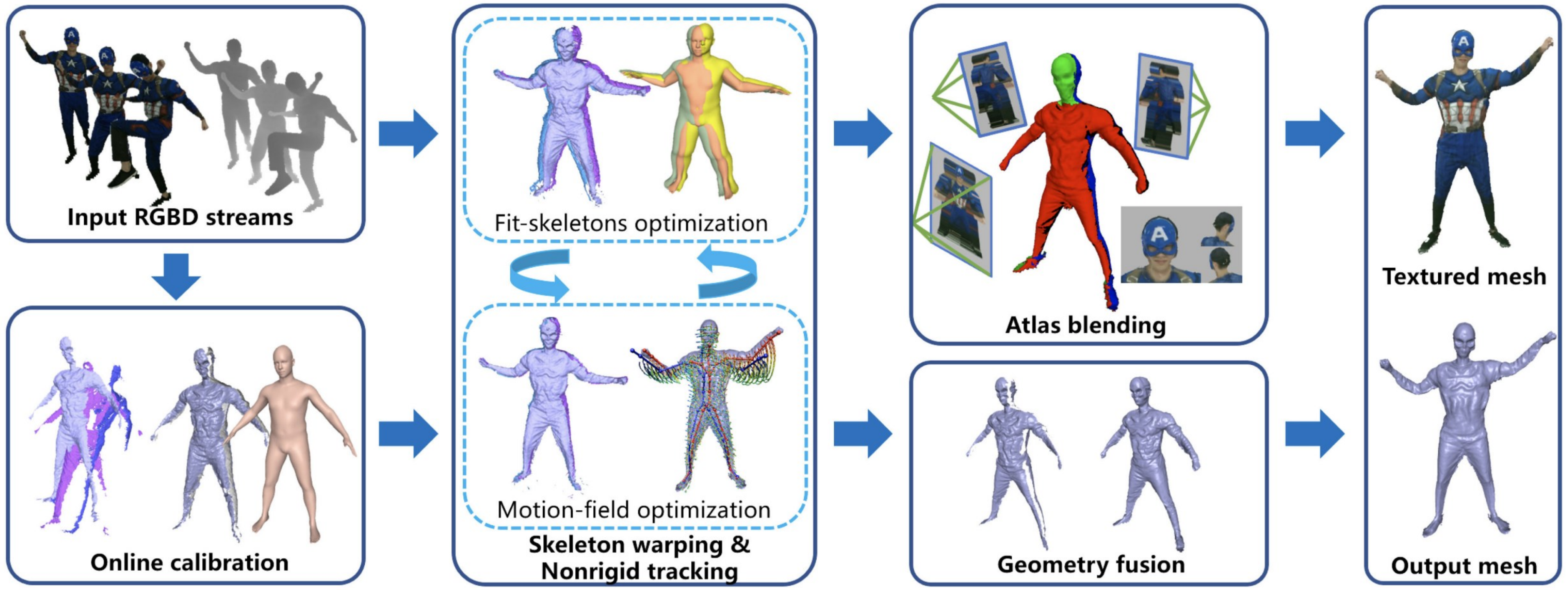
ByteDance字节跳动

2.1 Human Volumetric Capture

a. UnstructuredFusion: Realtime 4D Geometry and Texture Reconstruction using Commercial RGBD Cameras (TPAMI’20)

Texturing | **Dynamic atlas-based texture updating**

Designed a dynamic atlas texturing scheme to warp and update textures on fused geometry for high-quality real-time appearance reconstruction.



Asynchronous, unstructured, movable and real-time system

Tracking | **Skeleton warping-based non-rigid alignment**

Developed a non-rigid tracking scheme using skeleton warping for unstructured multiview depth alignment, supporting both online calibration and tracking.

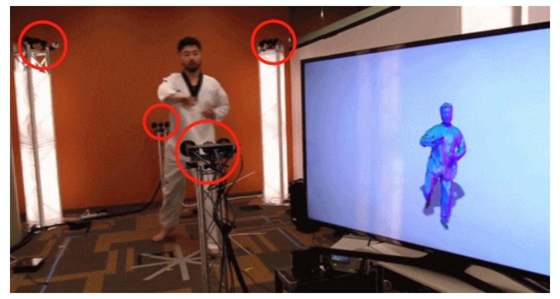
ByteDance字节跳动

2.1 传统方案：体积捕获

2.1 Human Volumetric Capture

Challenges#1

Multiple cameras rely on fixed spatial positions, making calibration and synchronization cumbersome.



8 custom RGBD cameras consisting of 24 individual cameras.

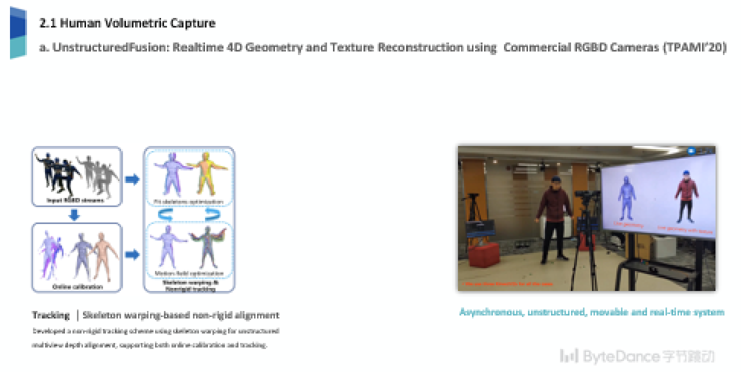
Challenges#2

The performer needs to rotate fully, and it is not robust for fast or complex motion tracking.

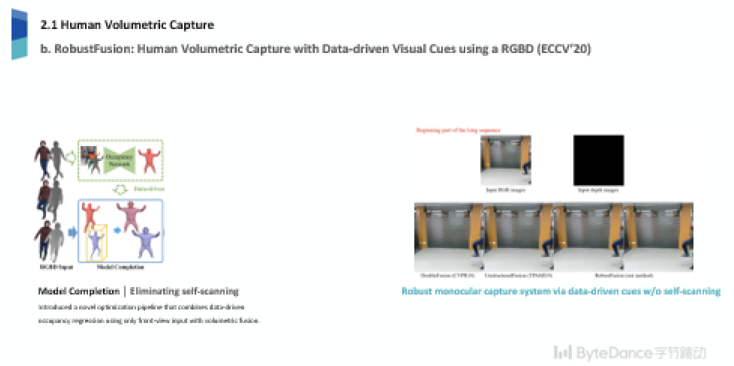
Challenges#3

It is challenging to robustly reconstruct human-object interaction scenes with complex motions.

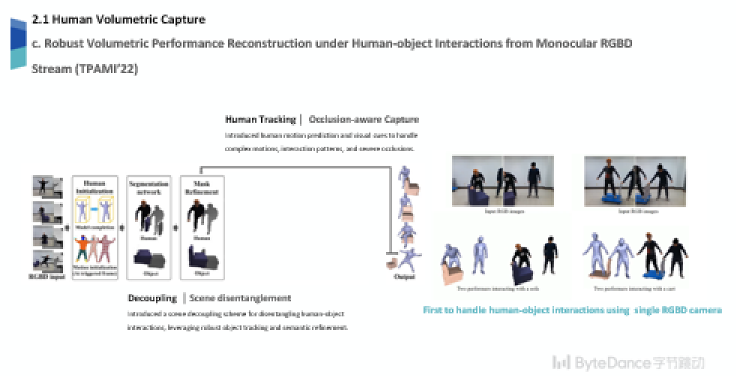
Structured & inflexible



Self-scanning at single-view



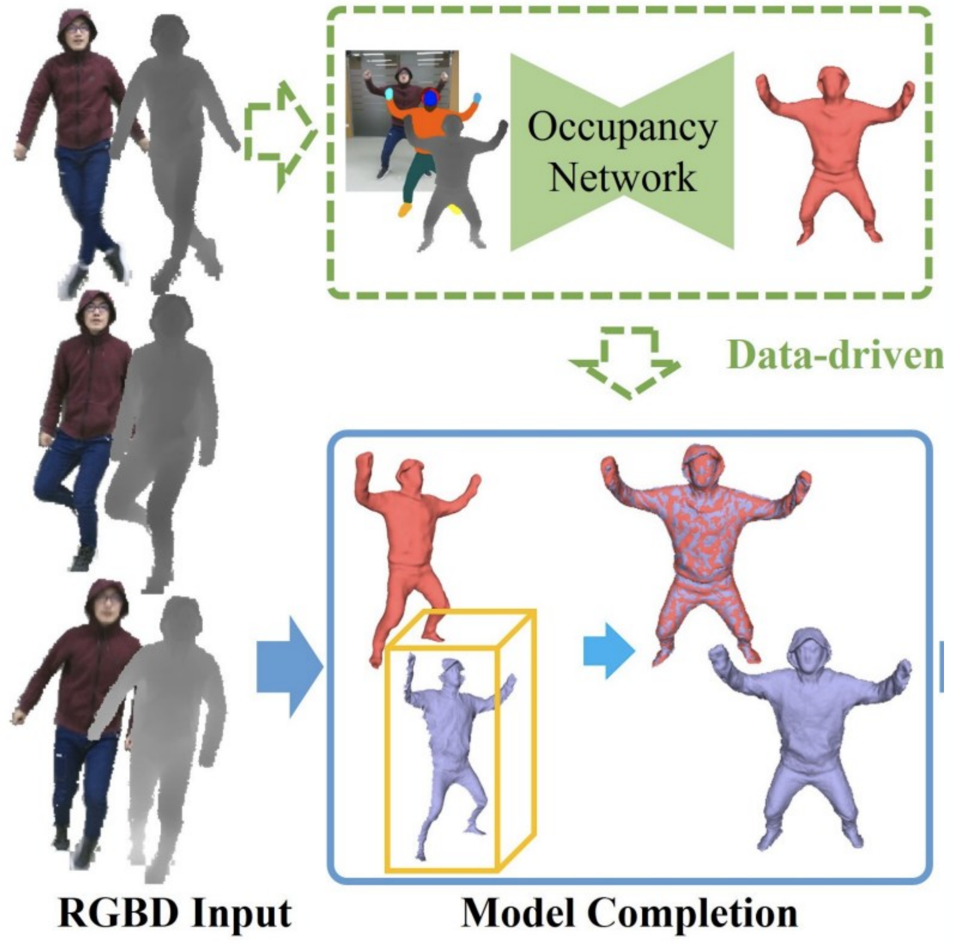
Human-object interaction

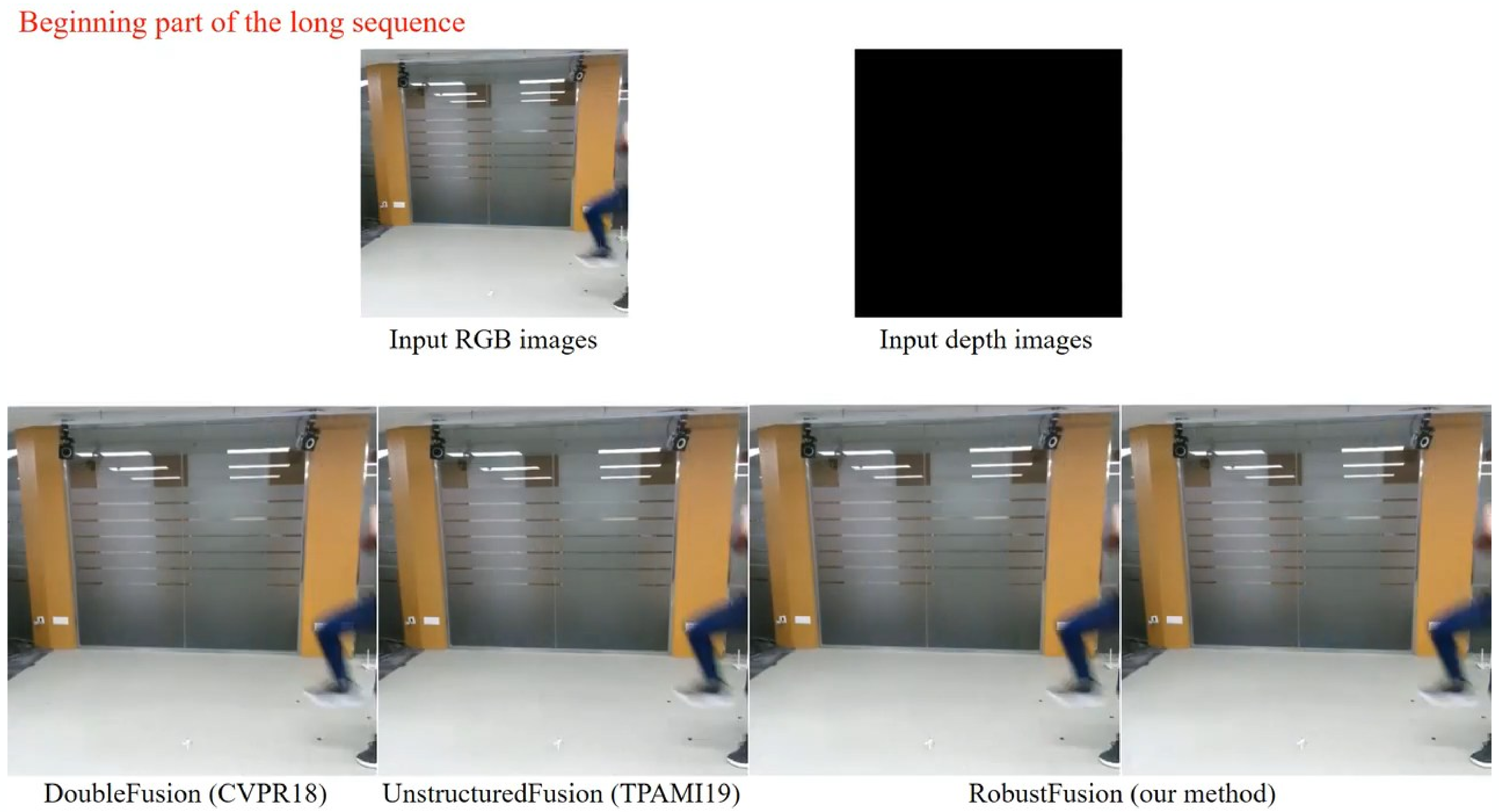


ByteDance字节跳动

2.1 Human Volumetric Capture

b. RobustFusion: Human Volumetric Capture with Data-driven Visual Cues using a RGBD (ECCV'20)





Model Completion | Eliminating self-scanning

Introduced a novel optimization pipeline that combines data-driven occupancy regression using only front-view input with volumetric fusion.

Robust monocular capture system via data-driven cues w/o self-scanning

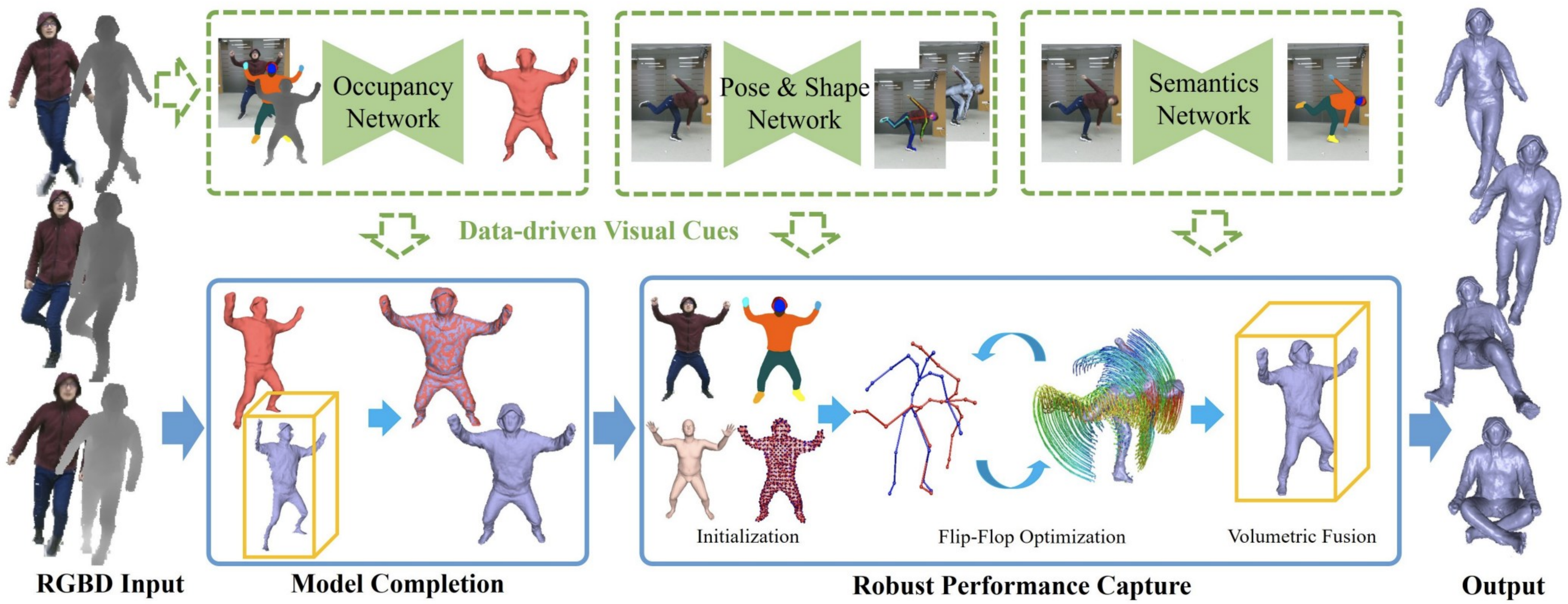
ByteDance字节跳动

2.1 Human Volumetric Capture

b. RobustFusion: Human Volumetric Capture with Data-driven Visual Cues using a RGBD (ECCV'20)

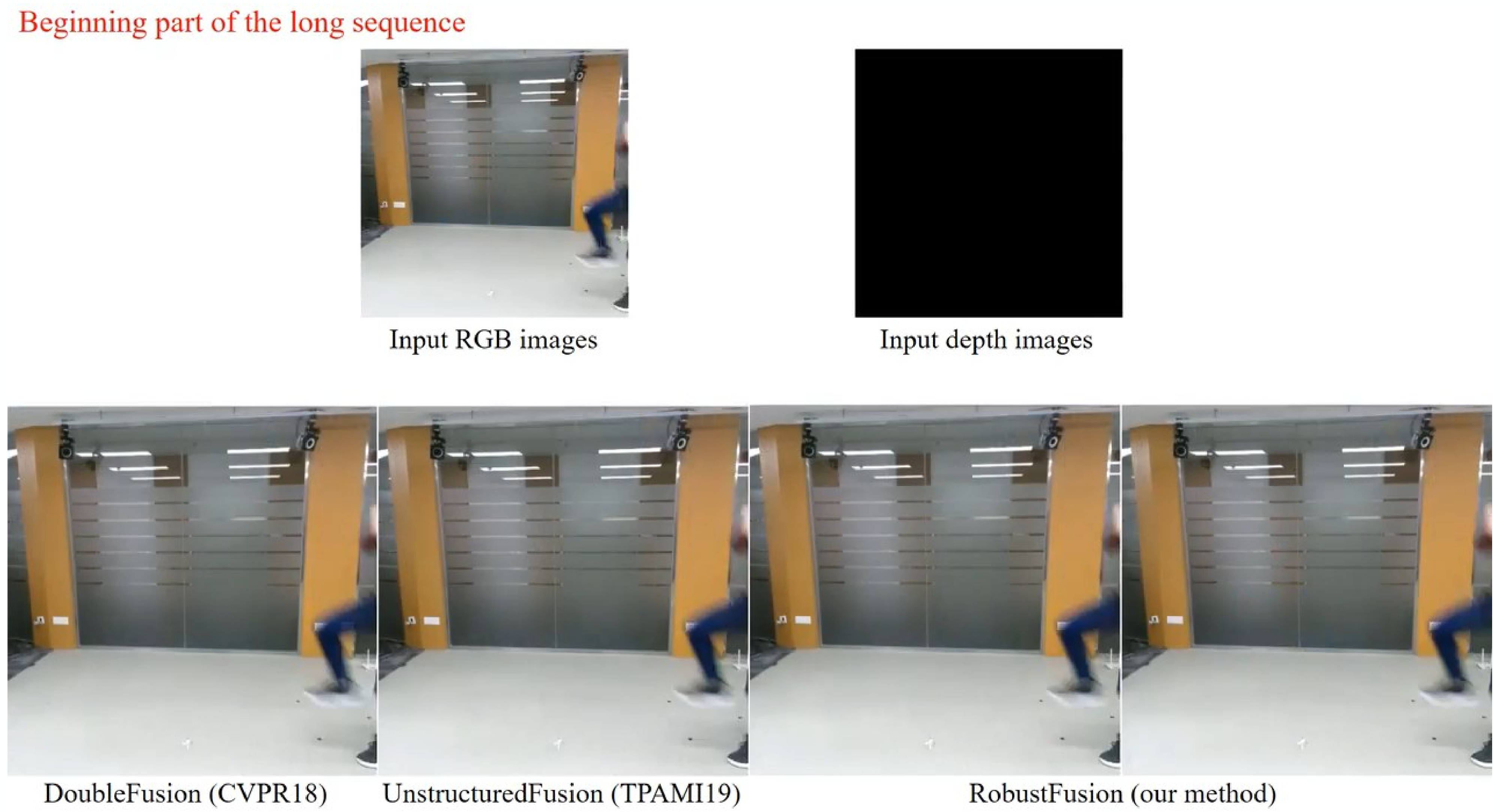
Tracking | Robust Performance Capture

Incorporated human pose, shape, and parsing priors to enable the handling of challenging human motions with reinitialization ability.



Model Completion | Eliminating self-scanning

Introduced a novel optimization pipeline that combines data-driven occupancy regression using only front-view input with volumetric fusion.



Robust monocular capture system via data-driven cues w/o self-scanning

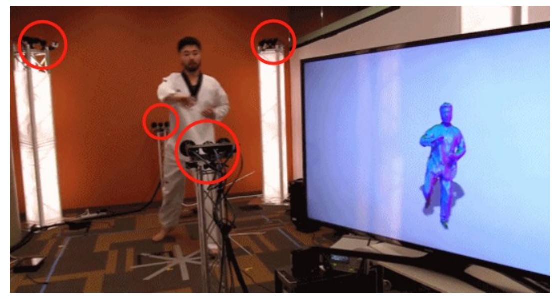
ByteDance字节跳动

2.1 传统方案：体积捕获

2.1 Human Volumetric Capture

Challenges#1

Multiple cameras rely on fixed spatial positions, making calibration and synchronization cumbersome.



8 custom RGBD cameras consisting of 24 individual cameras.

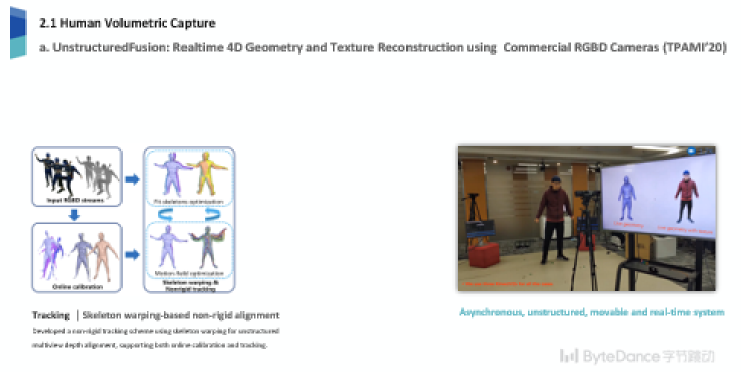
Challenges#2

The performer needs to rotate fully, and it is not robust for fast or complex motion tracking.

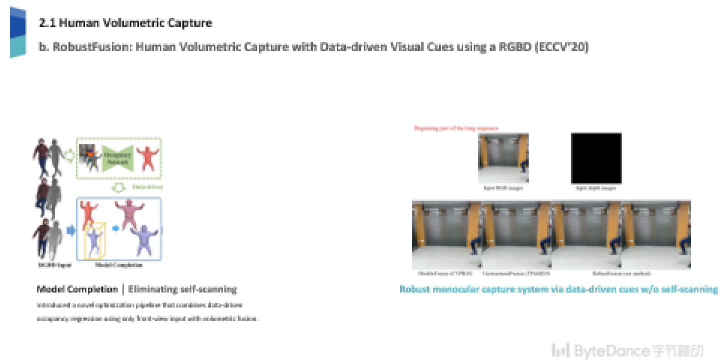
Challenges#3

It is challenging to robustly reconstruct human-object interaction scenes with complex motions.

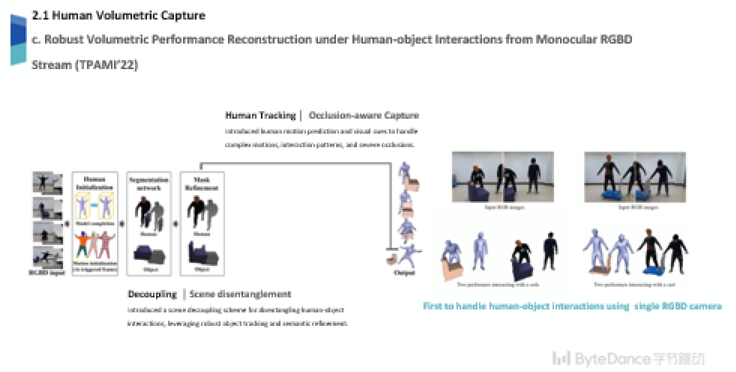
Structured & inflexible



Self-scanning at single-view



Human-object interaction



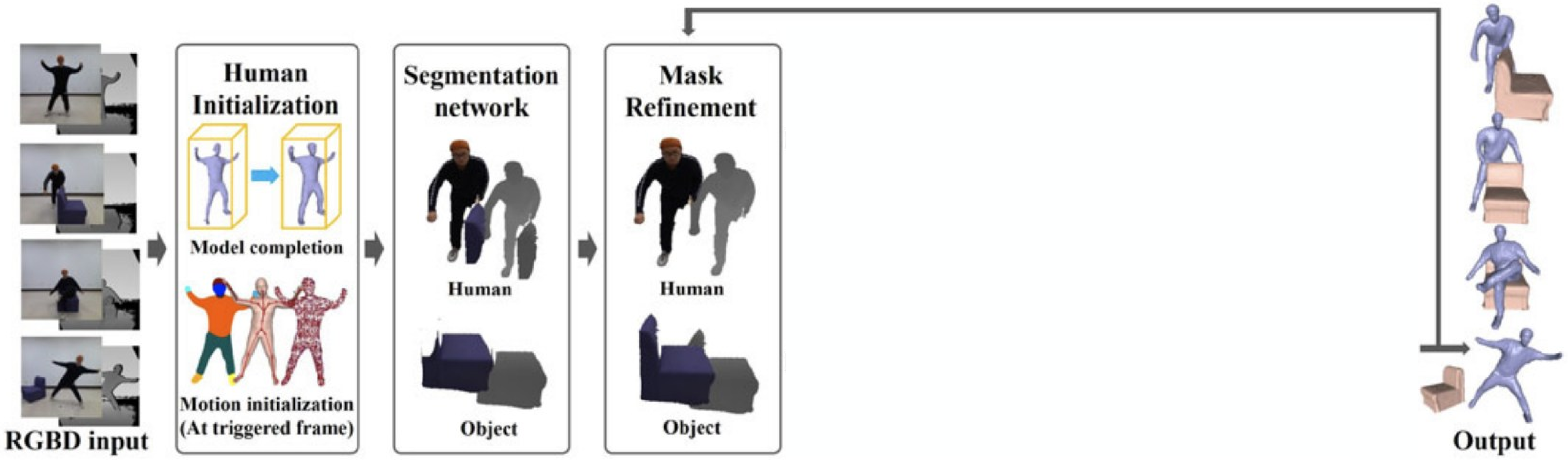
ByteDance字节跳动

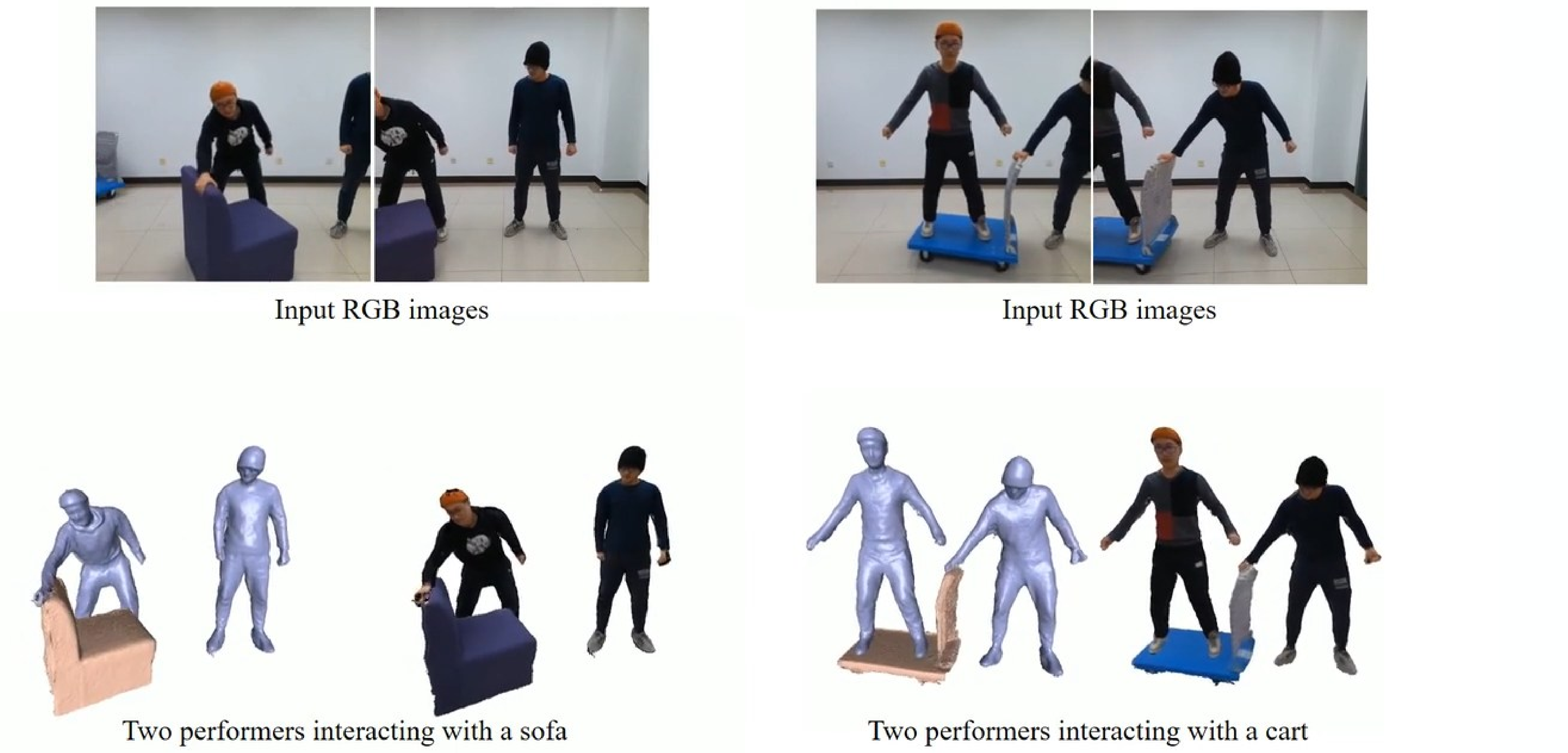
2.1 Human Volumetric Capture

c. Robust Volumetric Performance Reconstruction under Human-object Interactions from Monocular RGBD Stream (TPAMI'22)

Human Tracking | **Occlusion-aware Capture**

Introduced human motion prediction and visual cues to handle complex motions, interaction patterns, and severe occlusions.





Decoupling | **Scene disentanglement**

Introduced a scene decoupling scheme for disentangling human-object interactions, leveraging robust object tracking and semantic refinement.

First to handle human-object interactions using single RGBD camera

ByteDance字节跳动

2. 动态重建：从传统方案到神经渲染

2. Dynamic Reconstruction： From Volumetric Capture to Neural Rendering

2.1 传统方案：体积捕获

2.1 Human Volumetric Capture

2.2 隐式重建和神经渲染

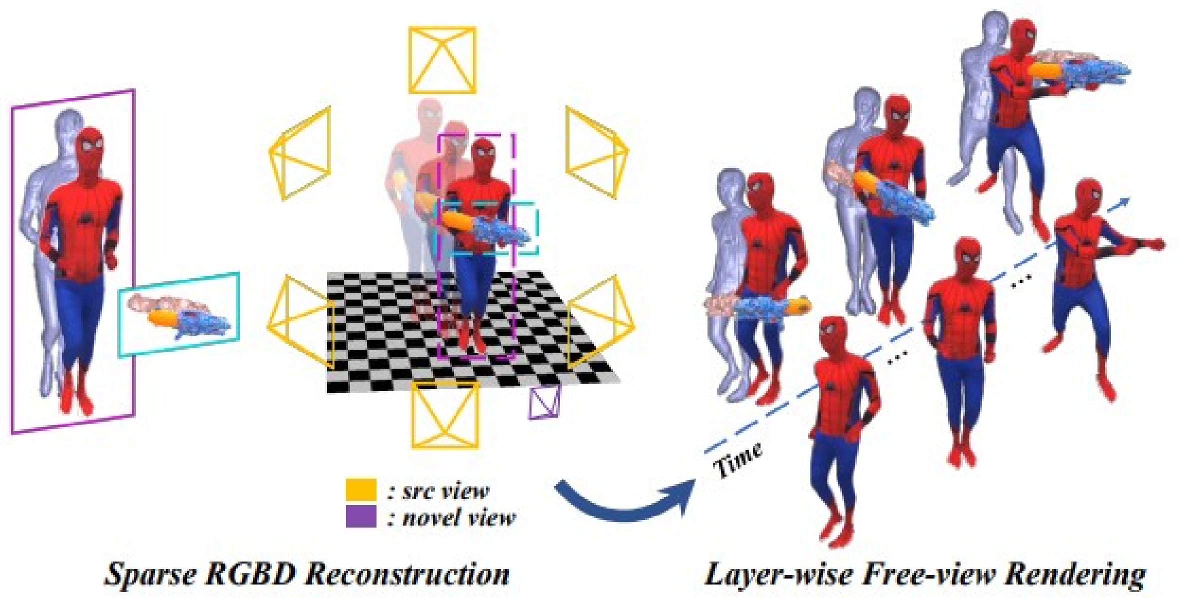
2.2 Human Neural Rendering

2.2 隐式重建和神经渲染

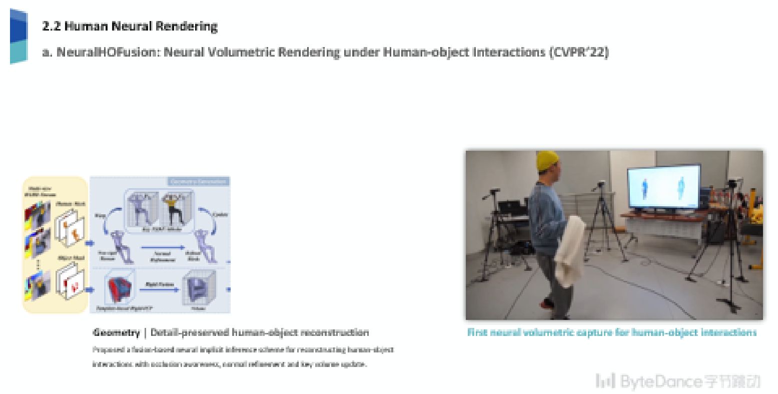
2.2 Human Neural Rendering

Challenges#1

The rendering of textures in explicit methods lacks clarity and cannot handle topology changes.

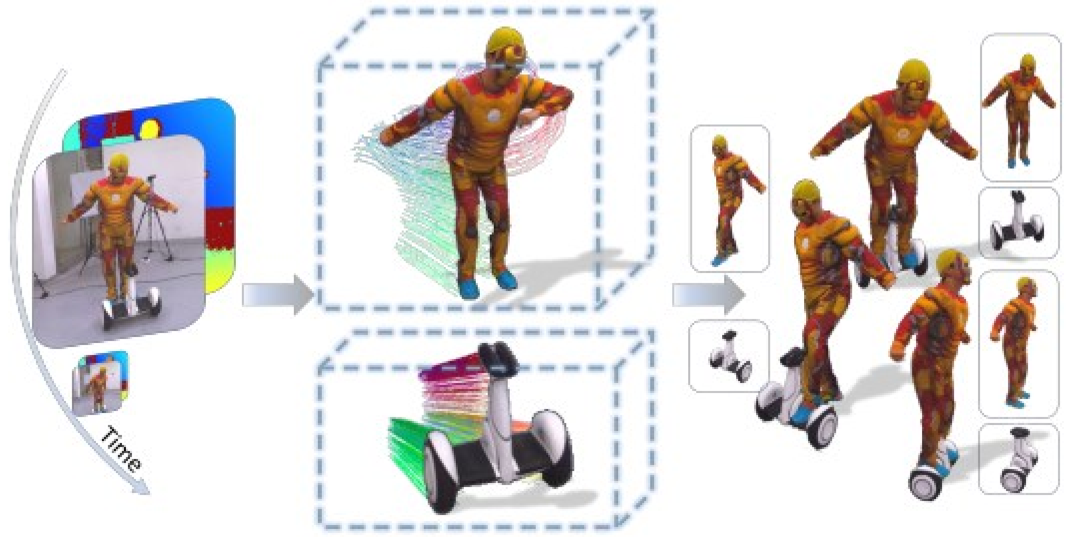


Topology changes & clearer textures

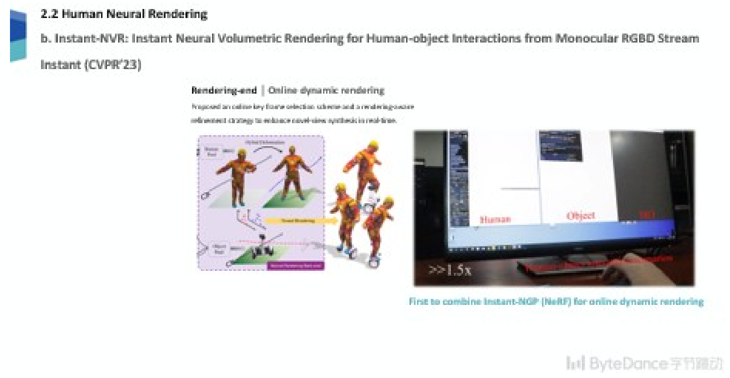


Challenges#2

On-the-fly neural rendering for dynamic human-object interactions in monocular settings is hindered by the need for slow NeRF training.



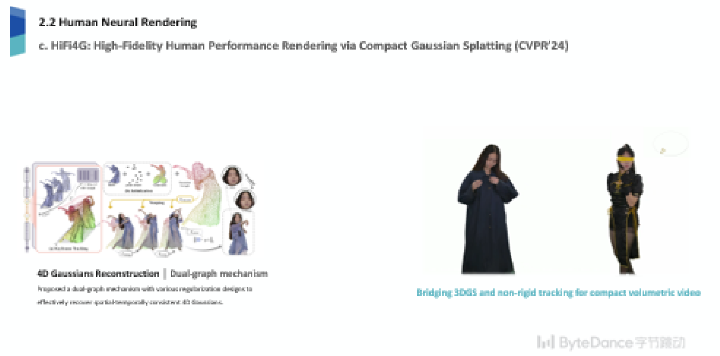
On-the-fly dynamic rendering



Challenges#3

Handling long motions with high quality and memory efficiency remains challenging.

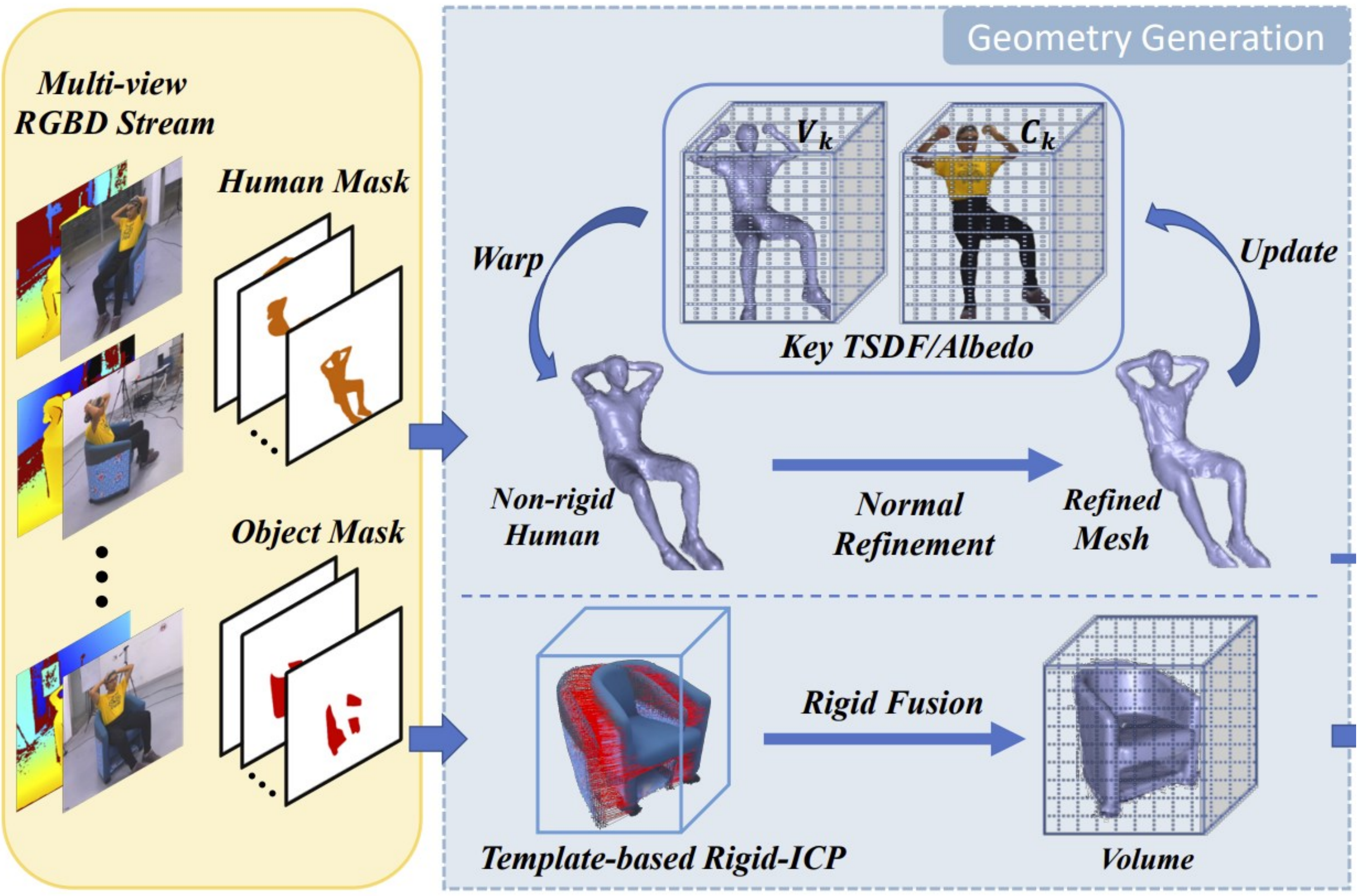
Human-object interaction



ByteDance字节跳动

2.2 Human Neural Rendering

a. NeuralHOFusion: Neural Volumetric Rendering under Human-object Interactions (CVPR'22)



Geometry | Detail-preserved human-object reconstruction

Proposed a fusion-based neural implicit inference scheme for reconstructing human-object interactions with occlusion awareness, normal refinement and key volume update.

First neural volumetric capture for human-object interactions

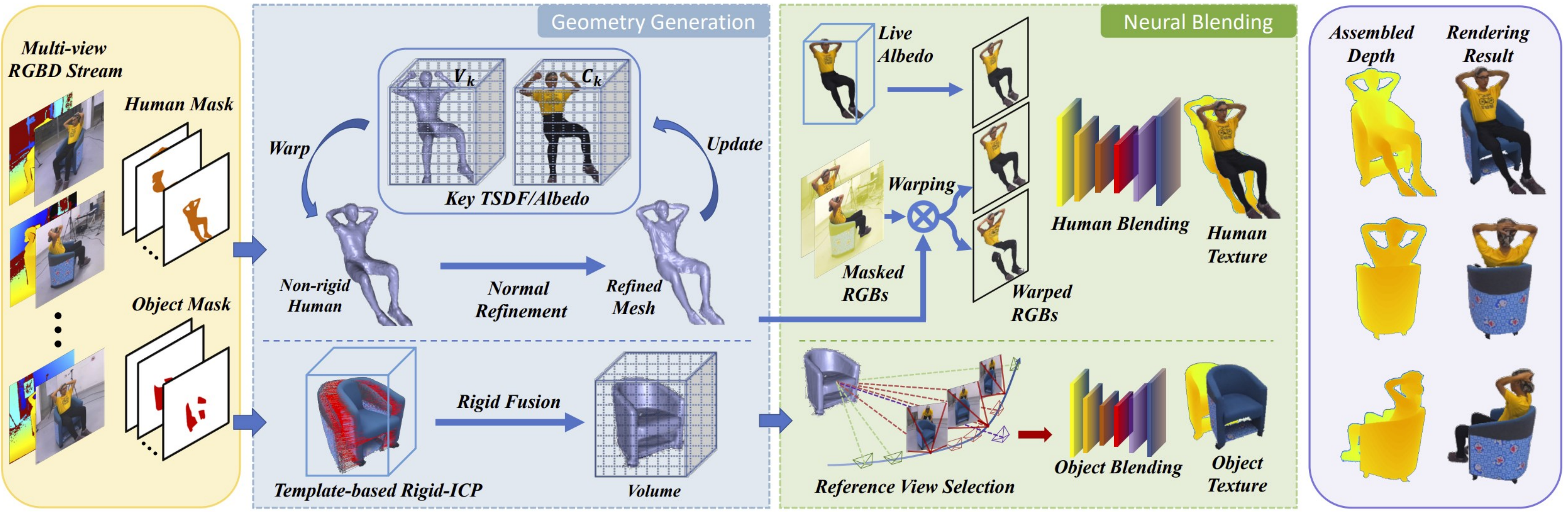
ByteDance字节跳动

2.2 Human Neural Rendering

a. NeuralHOFusion: Neural Volumetric Rendering under Human-object Interactions (CVPR’22)

Rendering | **Layer-wise blending-based neural rendering**

Introduced a layer-wise neural rendering approach, integrating volumetric and image-based rendering across spatial and temporal domains.



Geometry | **Detail-preserved human-object reconstruction**

Proposed a fusion-based neural implicit inference scheme for reconstructing human-object interactions with occlusion awareness, normal refinement and key volume update.

First neural volumetric capture for human-object interactions

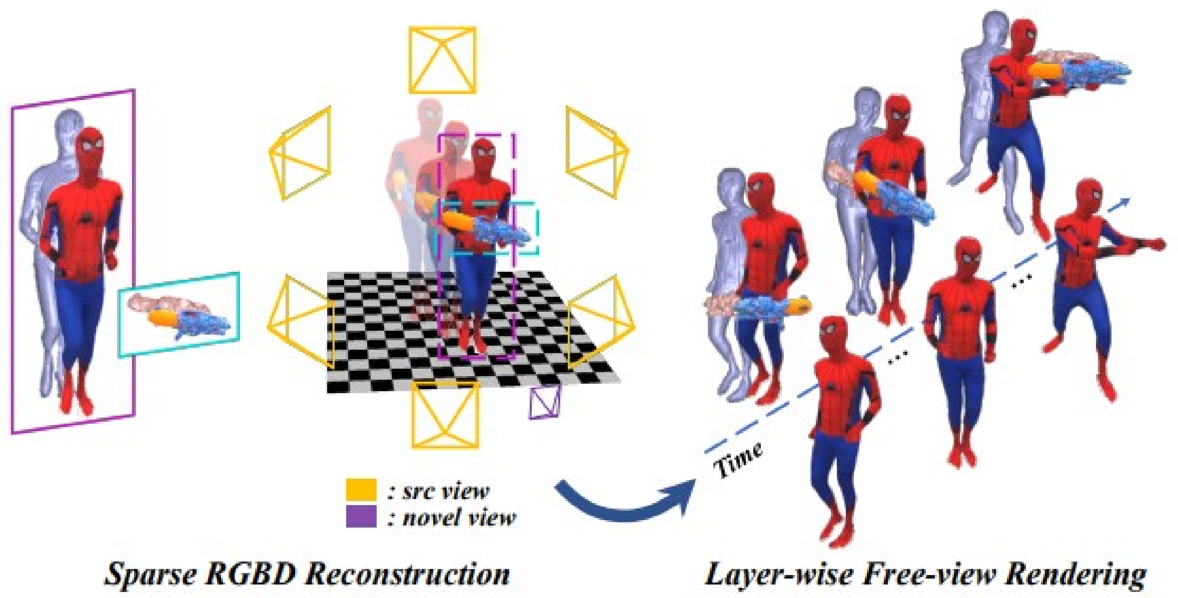
ByteDance字节跳动

2.2 隐式重建和神经渲染

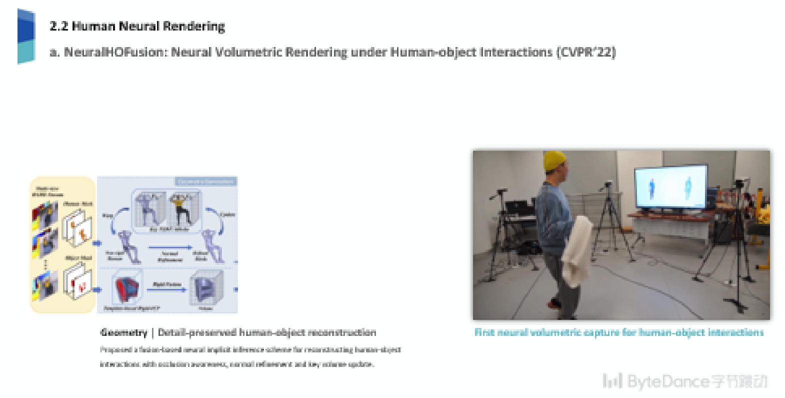
2.2 Human Neural Rendering

Challenges#1

The rendering of textures in explicit methods lacks clarity and cannot handle topology changes.

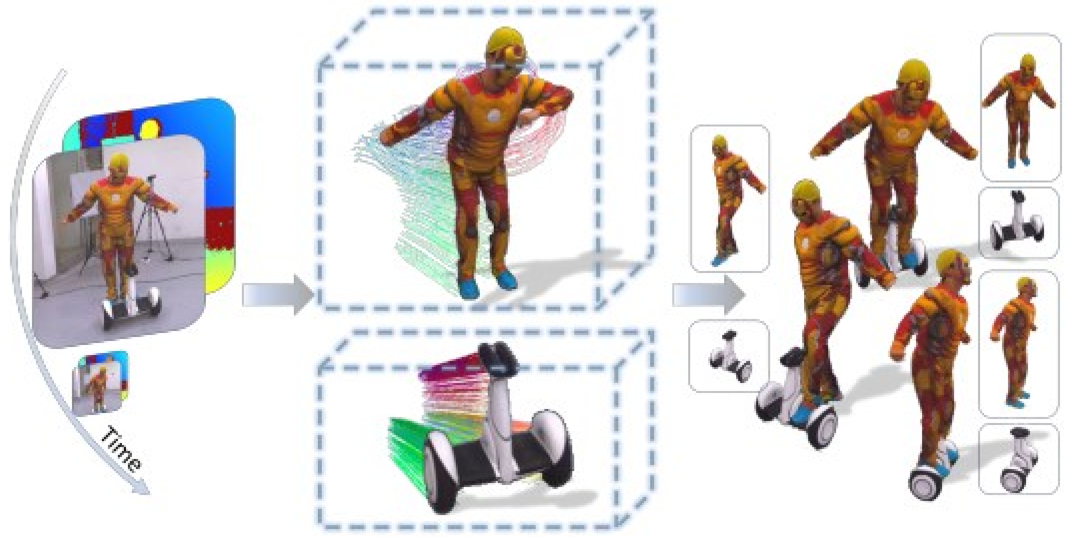


Topology changes & clearer textures

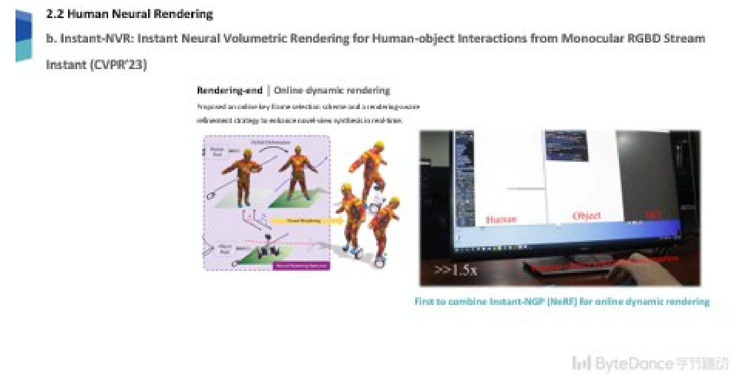


Challenges#2

On-the-fly neural rendering for dynamic human-object interactions in monocular settings is hindered by the need for slow NeRF training.



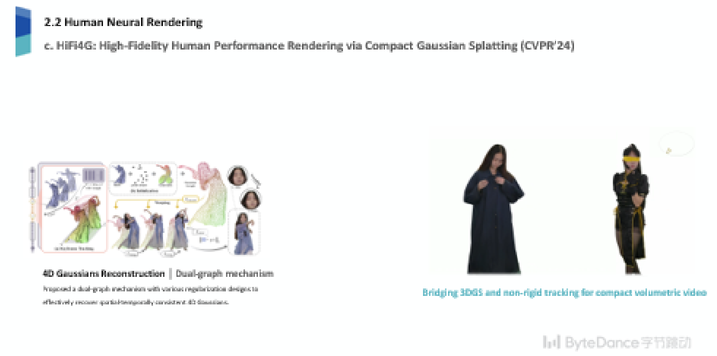
On-the-fly dynamic rendering



Challenges#3

Handling long motions with high quality and memory efficiency remains challenging.

Human-object interaction



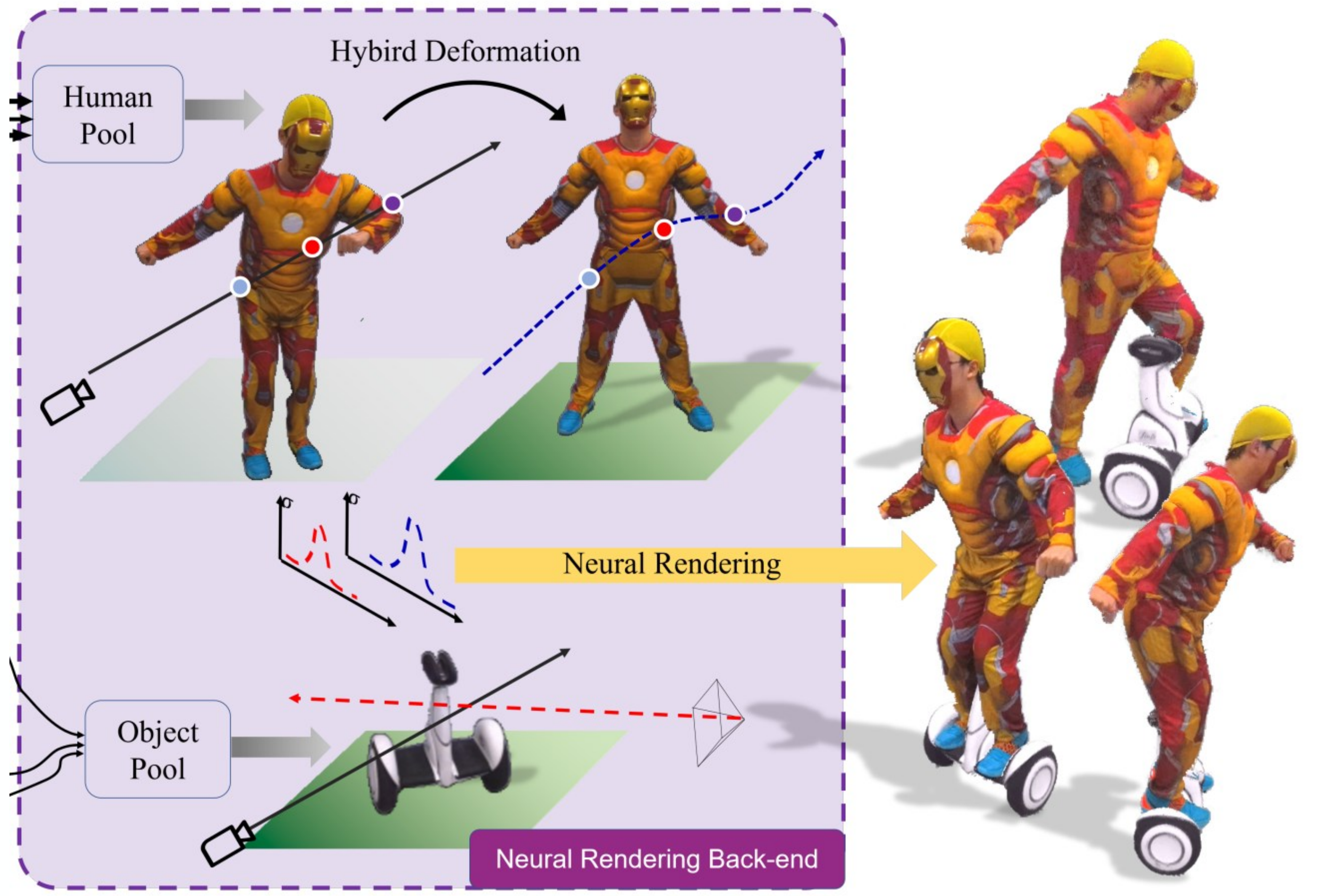
ByteDance字节跳动

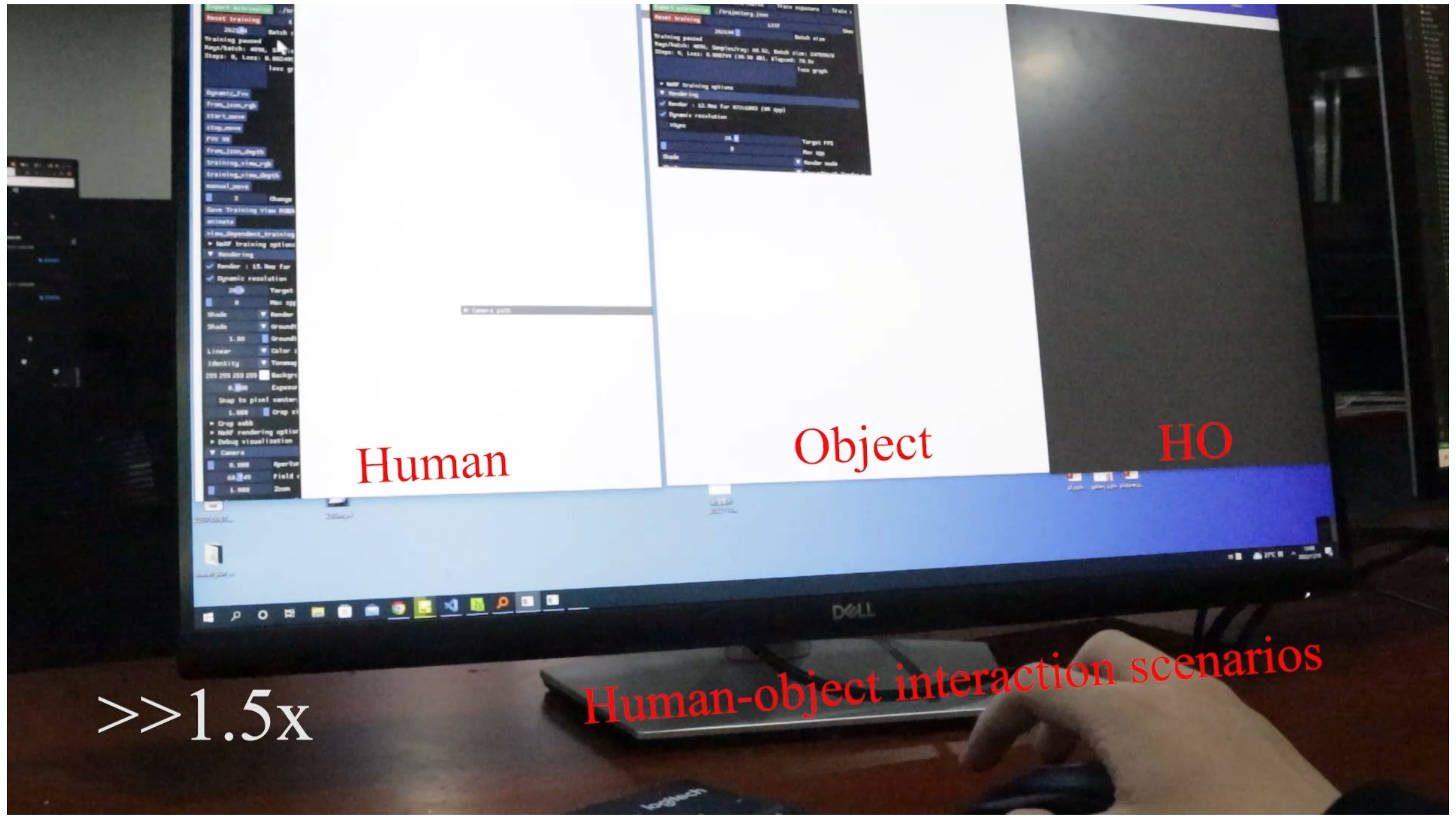
2.2 Human Neural Rendering

b. Instant-NVR: Instant Neural Volumetric Rendering for Human-object Interactions from Monocular RGBD Stream Instant (CVPR’23)

Rendering-end | Online dynamic rendering

Proposed an online key frame selection scheme and a rendering-aware refinement strategy to enhance novel-view synthesis in real-time.





First to combine Instant-NGP (NeRF) for online dynamic rendering

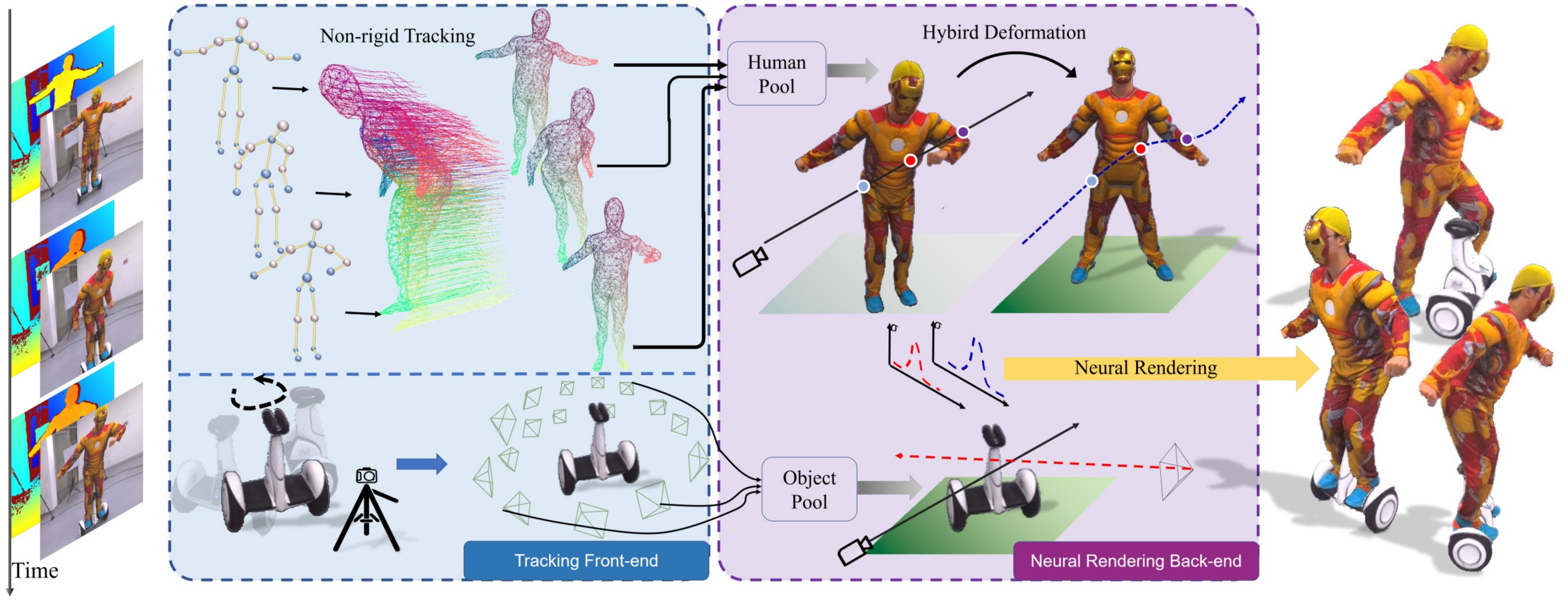
ByteDance字节跳动

2.2 Human Neural Rendering

b. Instant-NVR: Instant Neural Volumetric Rendering for Human-object Interactions from Monocular RGBD Stream Instant (CVPR'23)

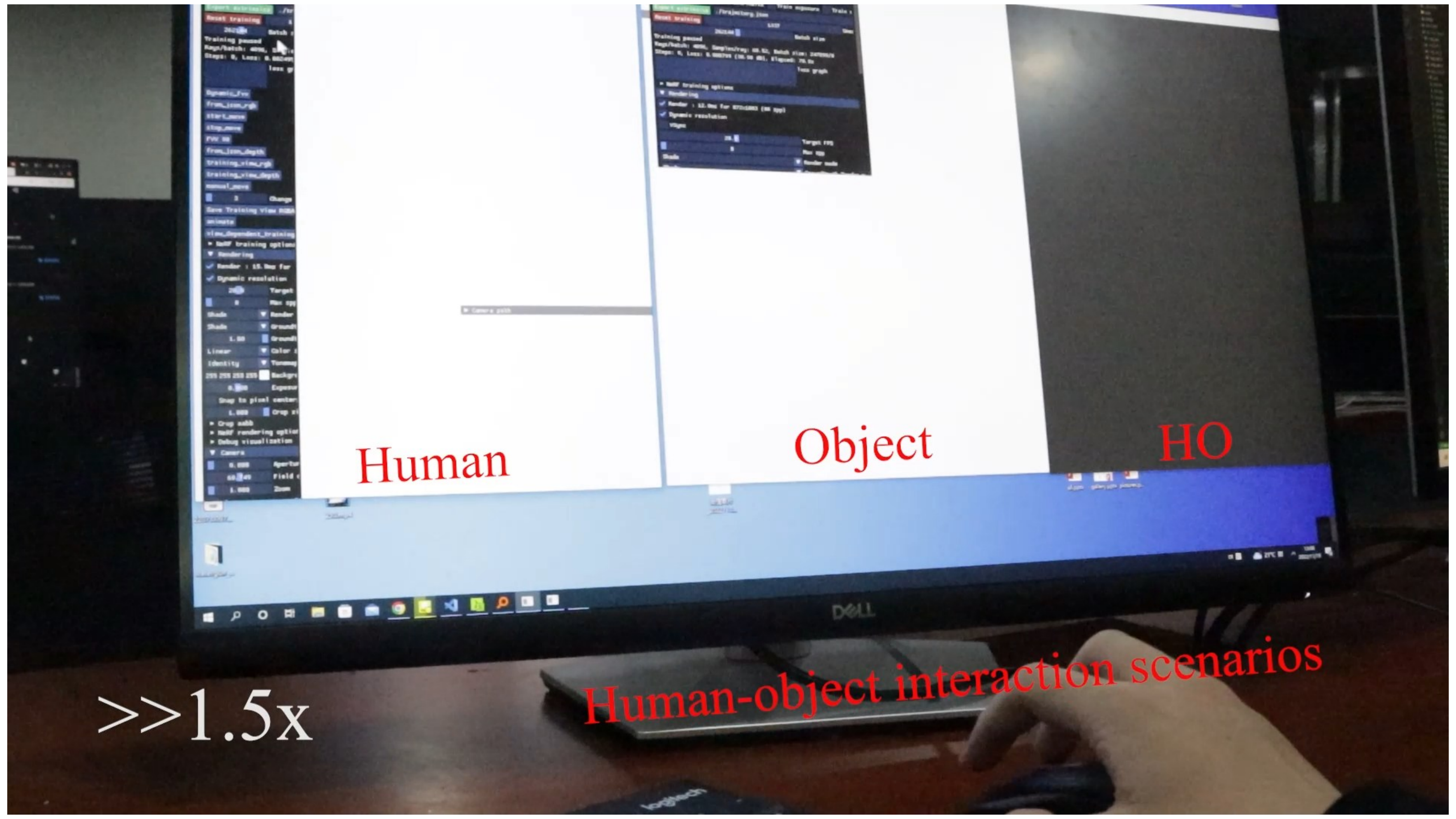
Rendering-end | Online dynamic rendering

Proposed an online key frame selection scheme and a rendering-aware refinement strategy to enhance novel-view synthesis in real-time.



Tracking-end | On-the-fly reconstruction

Introduced an on-the-fly reconstruction scheme for dynamic and static radiance fields, leveraging motion priors through a tracking-rendering mechanism.



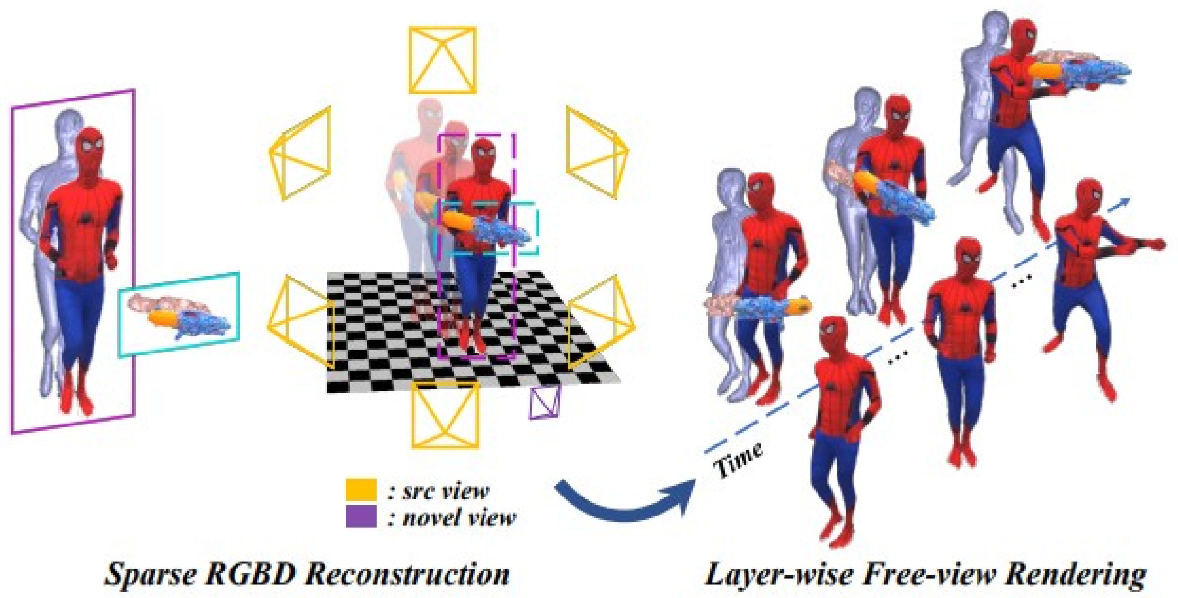
First to combine Instant-NGP (NeRF) for online dynamic rendering

2.2 隐式重建和神经渲染

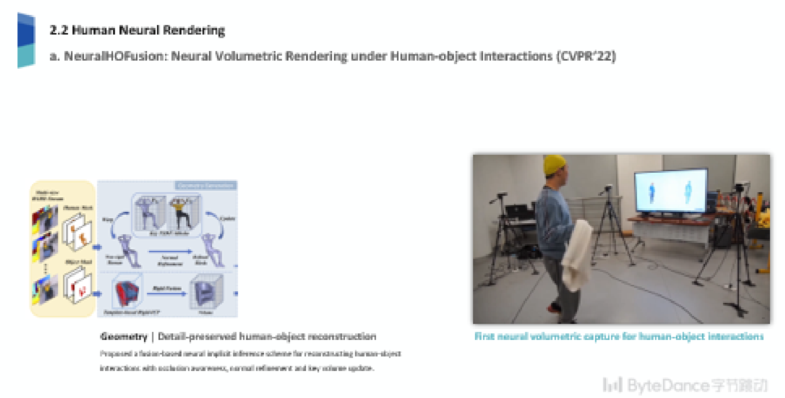
2.2 Human Neural Rendering

Challenges#1

The rendering of textures in explicit methods lacks clarity and cannot handle topology changes.

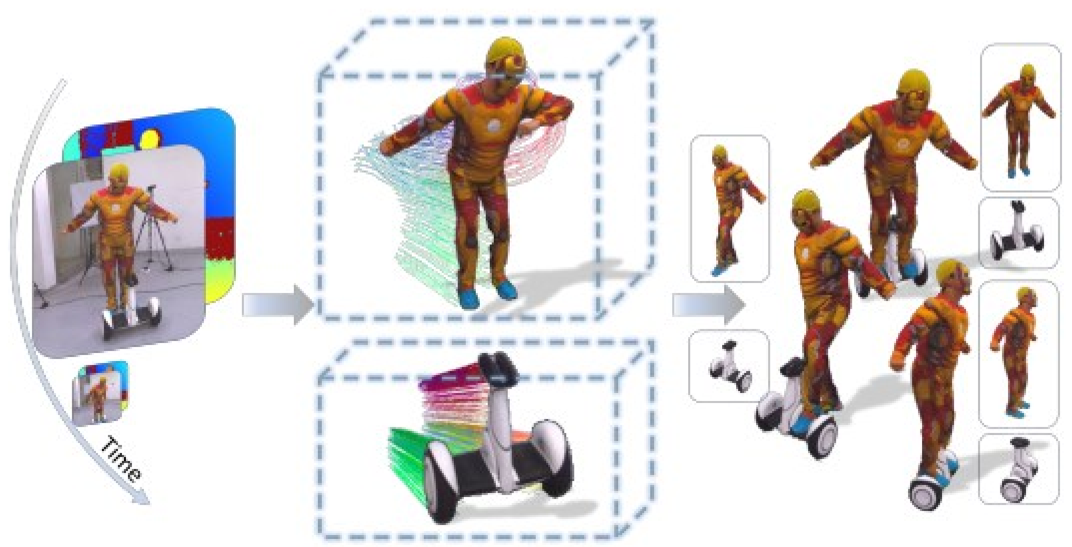


Topology changes & clearer textures

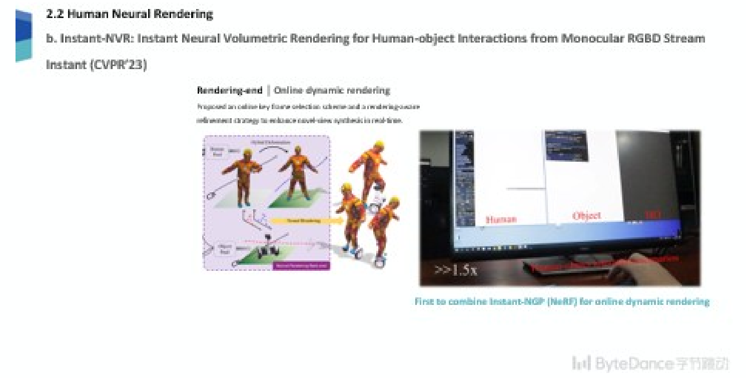


Challenges#2

On-the-fly neural rendering for dynamic human-object interactions in monocular settings is hindered by the need for slow NeRF training.



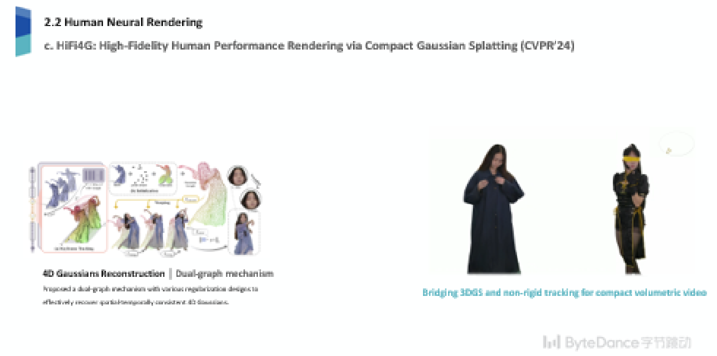
On-the-fly dynamic rendering



Challenges#3

Handling long motions with high quality and memory efficiency remains challenging.

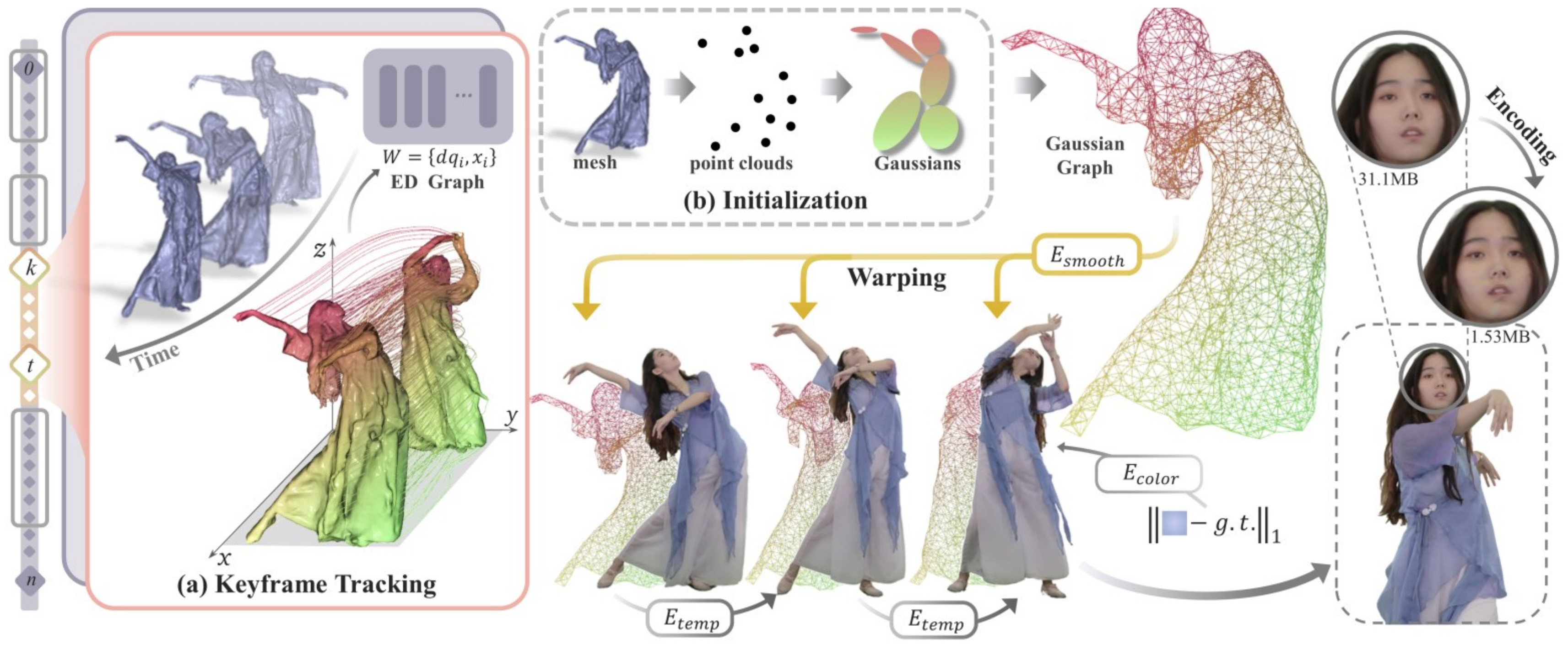
High quality & low storage



ByteDance字节跳动

2.2 Human Neural Rendering

c. HiFi4G: High-Fidelity Human Performance Rendering via Compact Gaussian Splatting (CVPR'24)



4D Gaussians Reconstruction | Dual-graph mechanism

Proposed a dual-graph mechanism with various regularization designs to effectively recover spatial-temporally consistent 4D Gaussians.

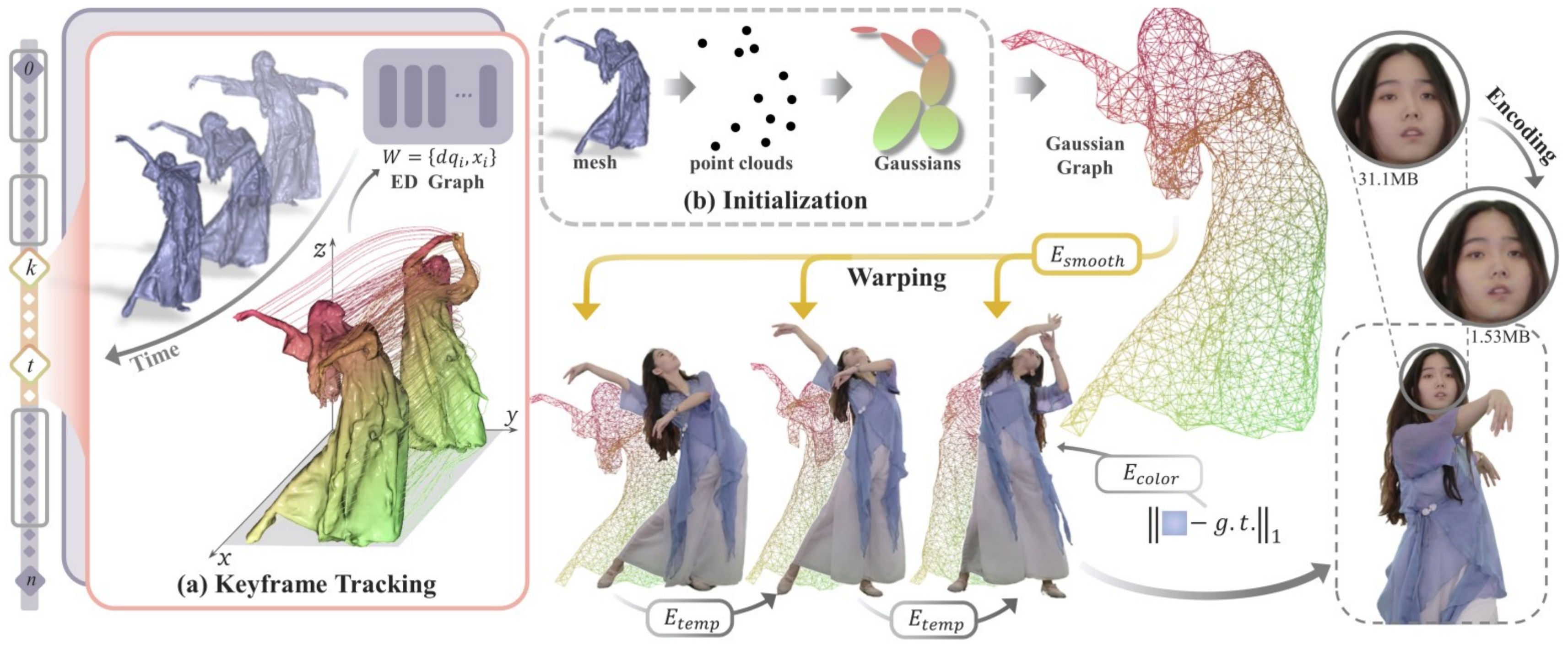
Bridging 3DGS and non-rigid tracking for compact volumetric video

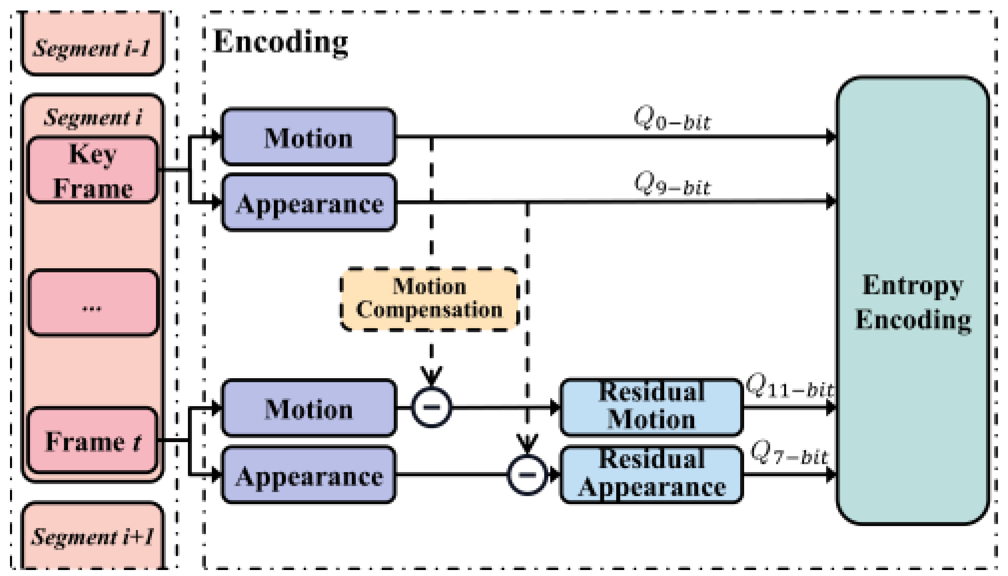
2.2 Human Neural Rendering

c. HiFi4G: High-Fidelity Human Performance Rendering via Compact Gaussian Splatting (CVPR'24)

Compact Representation | Compression scheme

Showcased a compression scheme, supporting immersive experience of human performance with low storage, even under various platforms such as VR devices.





4D Gaussians Reconstruction | Dual-graph mechanism

Proposed a dual-graph mechanism with various regularization designs to effectively recover spatial-temporally consistent 4D Gaussians.

Bridging 3DGS and non-rigid tracking for compact volumetric video

2. 动态重建：从传统方案到神经渲染

2. Dynamic Reconstruction：From Volumetric Capture to Neural Rendering

2.1 传统方案：体积捕获

2.1 Human Volumetric Capture

2.2 隐式重建和神经渲染

2.2 Human Neural Rendering

Optimization + Explicit 3D	
	学习化
Data-driven cues introducing	
	隐式化
Learning + Implicit 3D	

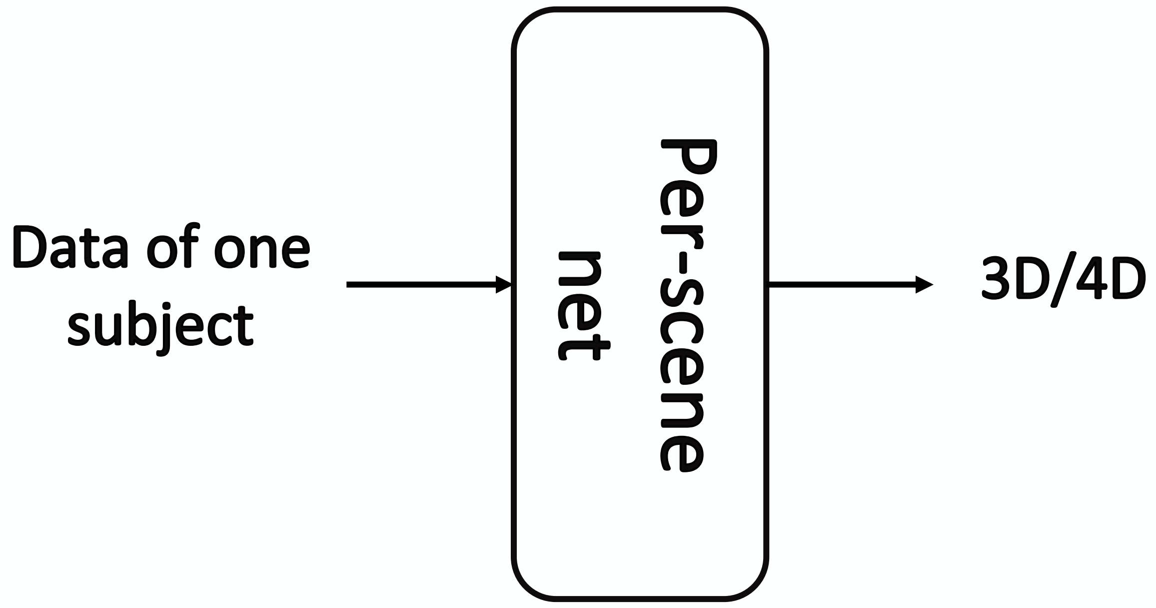
ByteDance字节跳动

3. 形象重建：从先验模型到3D生成

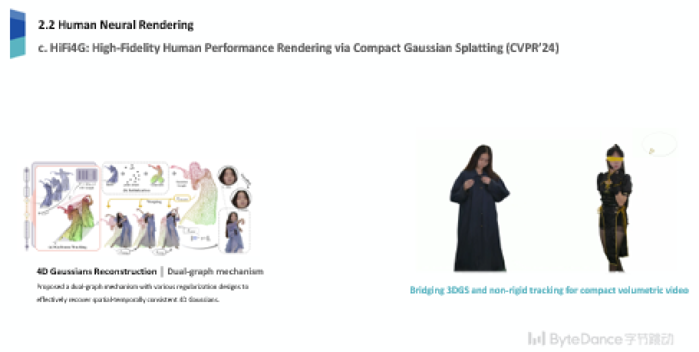
3. Avatar Reconstruction：From Prior model to 3D Generation

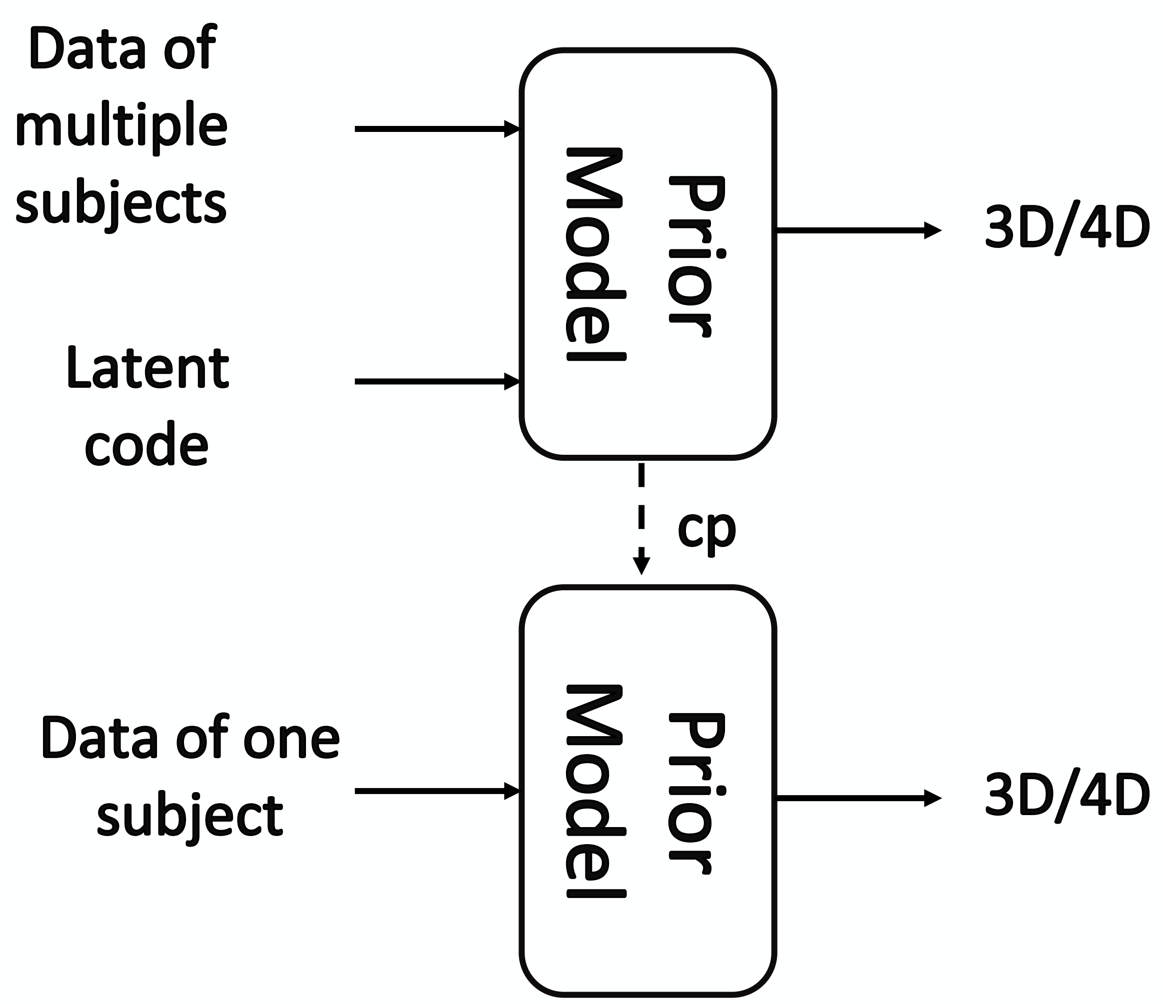
3.1 基于先验模型的形象重建

3.1 Animatable Avatar Creation

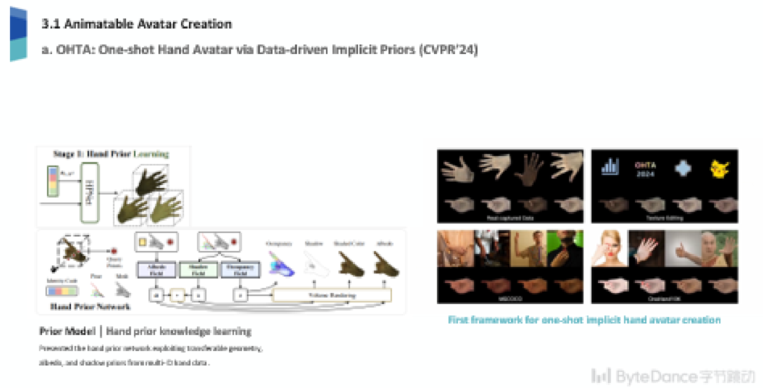


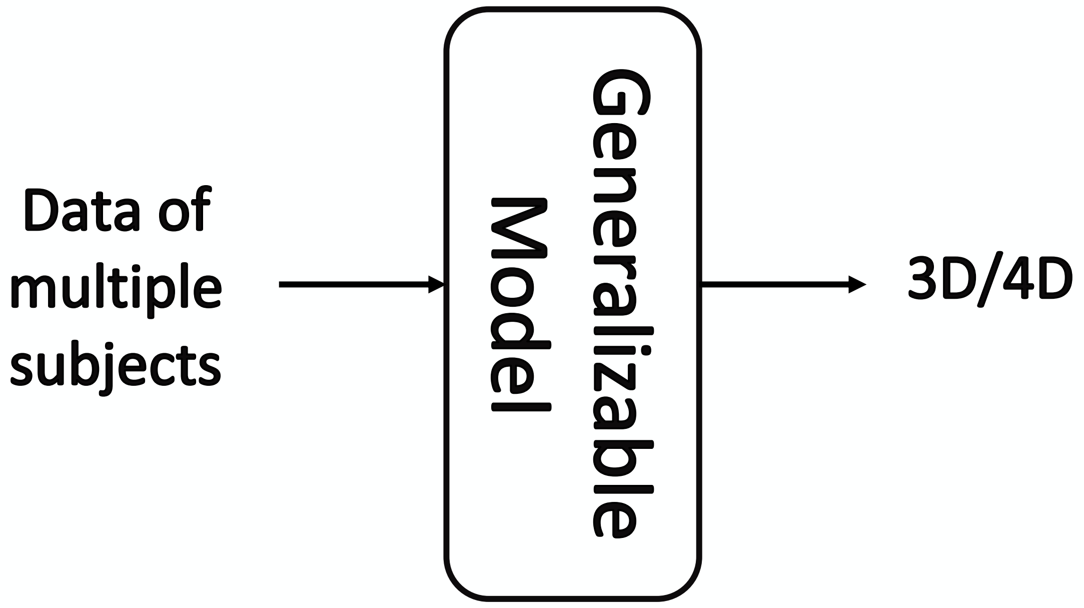
Per-scene training



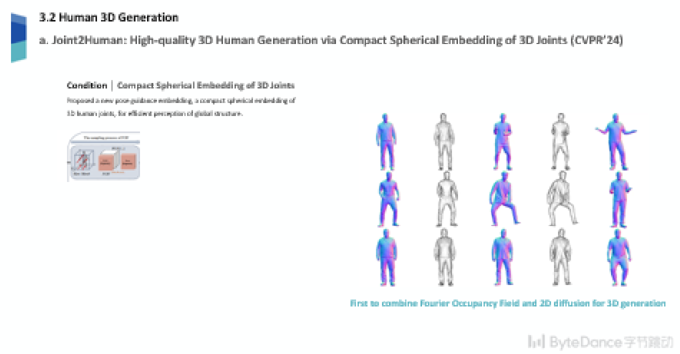


Prior training & tuning





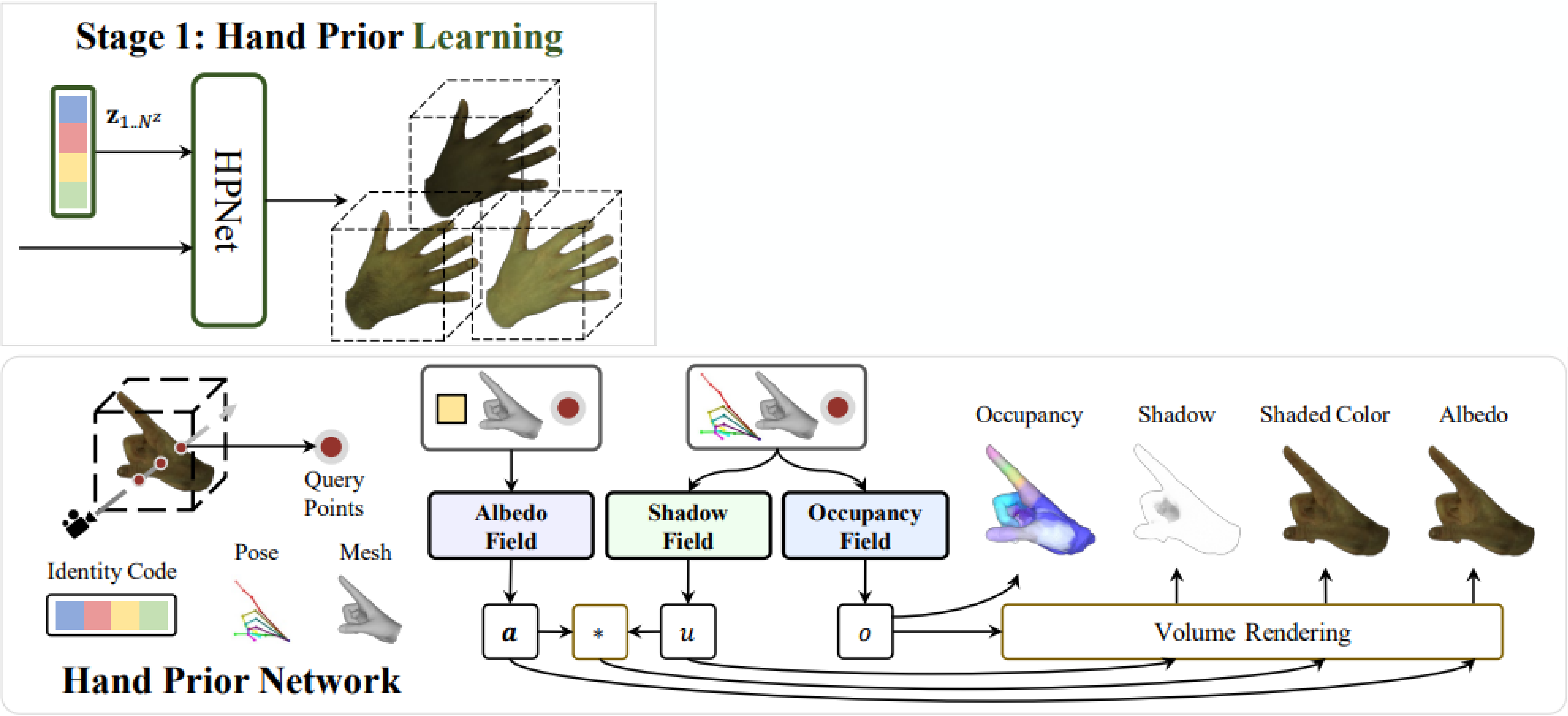
Generalizable training



ByteDance字节跳动

3.1 Animatable Avatar Creation

a. OHTA: One-shot Hand Avatar via Data-driven Implicit Priors (CVPR’24)



First framework for one-shot implicit hand avatar creation

Prior Model | Hand prior knowledge learning

Presented the hand prior network exploiting transferable geometry, albedo, and shadow priors from multi-ID hand data.

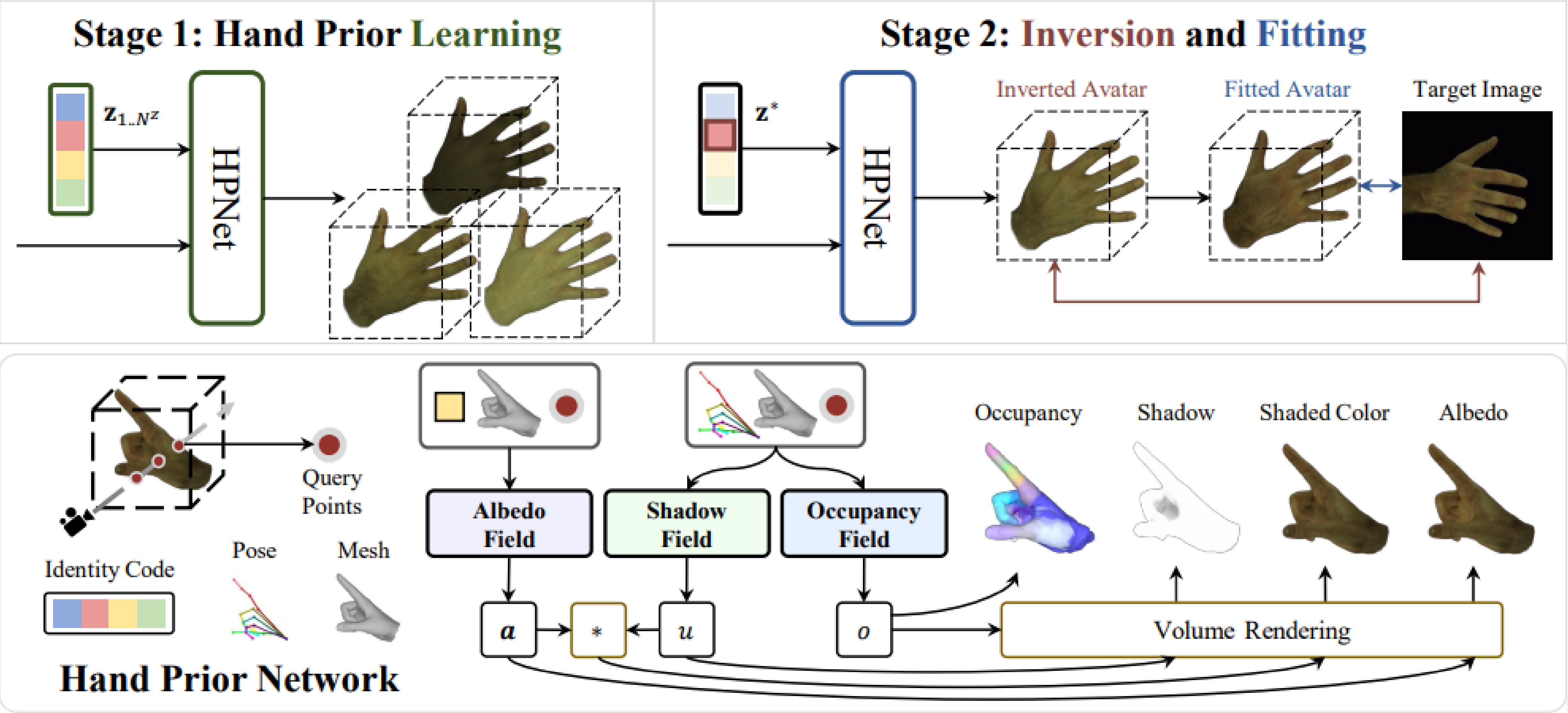
ByteDance字节跳动

3.1 Animatable Avatar Creation

a. OHTA: One-shot Hand Avatar via Data-driven Implicit Priors (CVPR’24)

Personalization | One-shot reconstruction

Achieved high-fidelity of one-shot hand avatar creation by leveraging inversion and fitting strategies, showcasing applications like text-to-avatar, editing, and interpolation.



First framework for one-shot implicit hand avatar creation

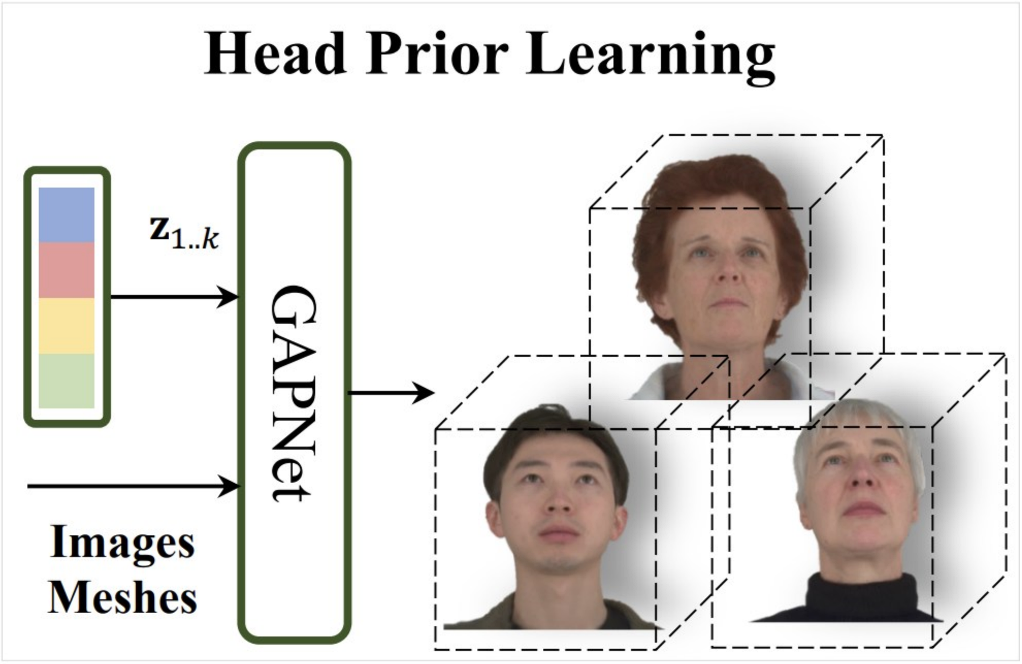
Prior Model | Hand prior knowledge learning

Presented the hand prior network exploiting transferable geometry, albedo, and shadow priors from multi-ID hand data.

ByteDance字节跳动

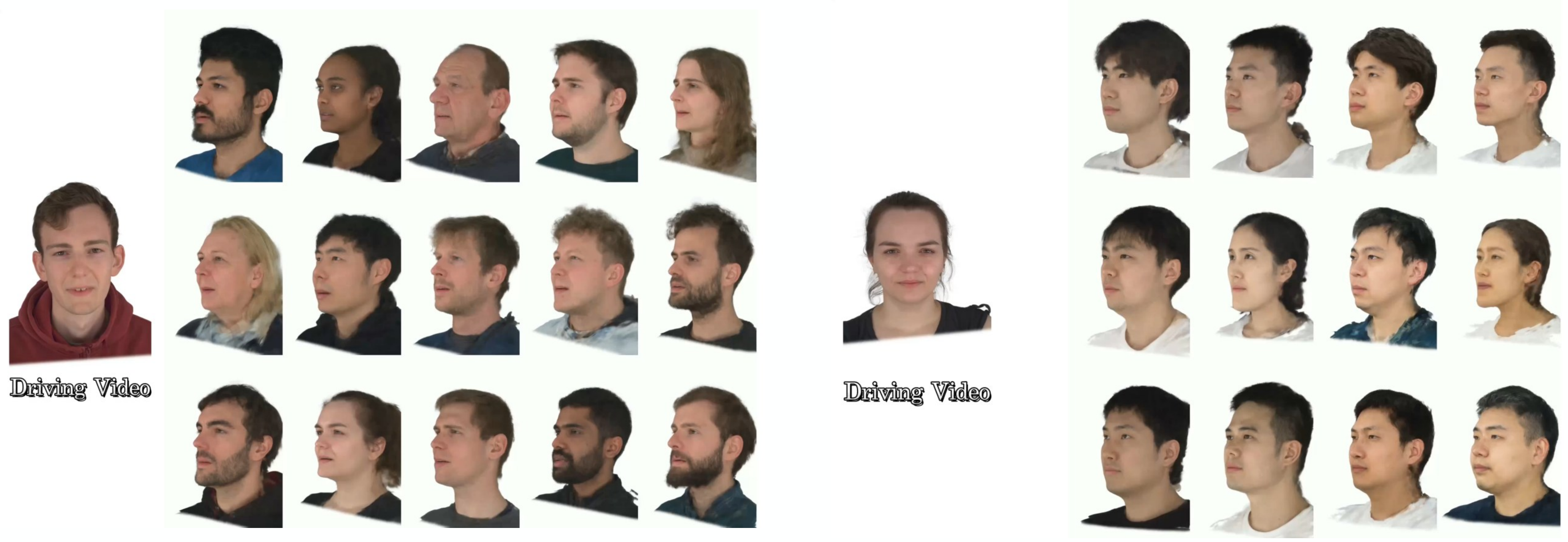
3.1 Animatable Avatar Creation

b. HeadGAP: Few-shot 3D Head Avatar via Generalizable Gaussian Priors Instant (3DV’25)



Prior Model | Auto-decoder framework

Presented auto-decoder designs that effectively utilize part-based dynamic Gaussian head priors trained from multi-ID, multi-view and multi-expression data.



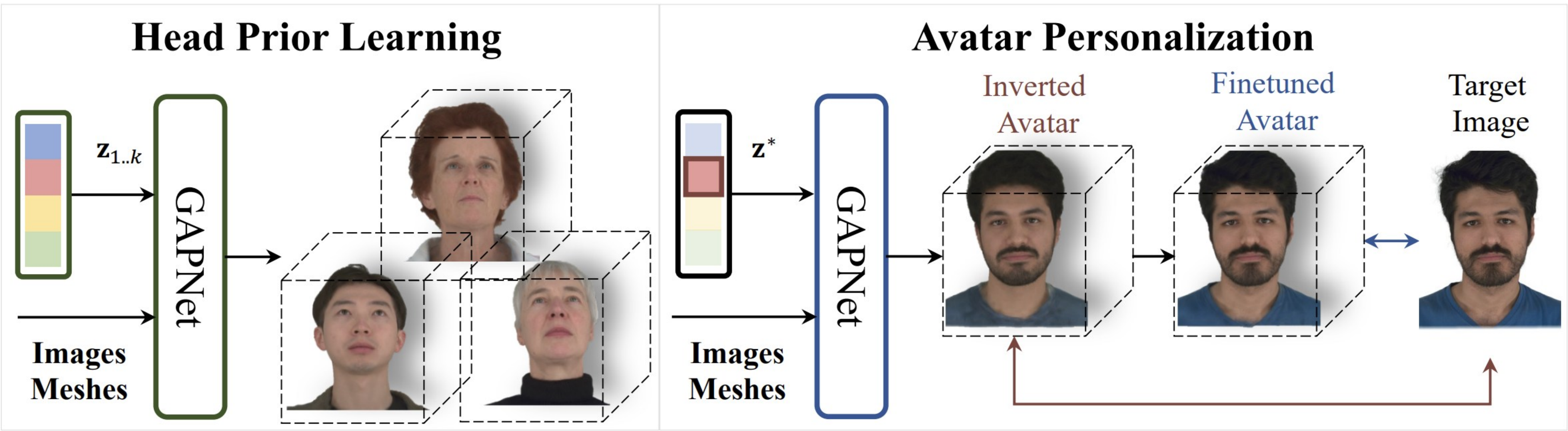
First to leverage generalizable 3DGS priors for few-shot head avatar creation

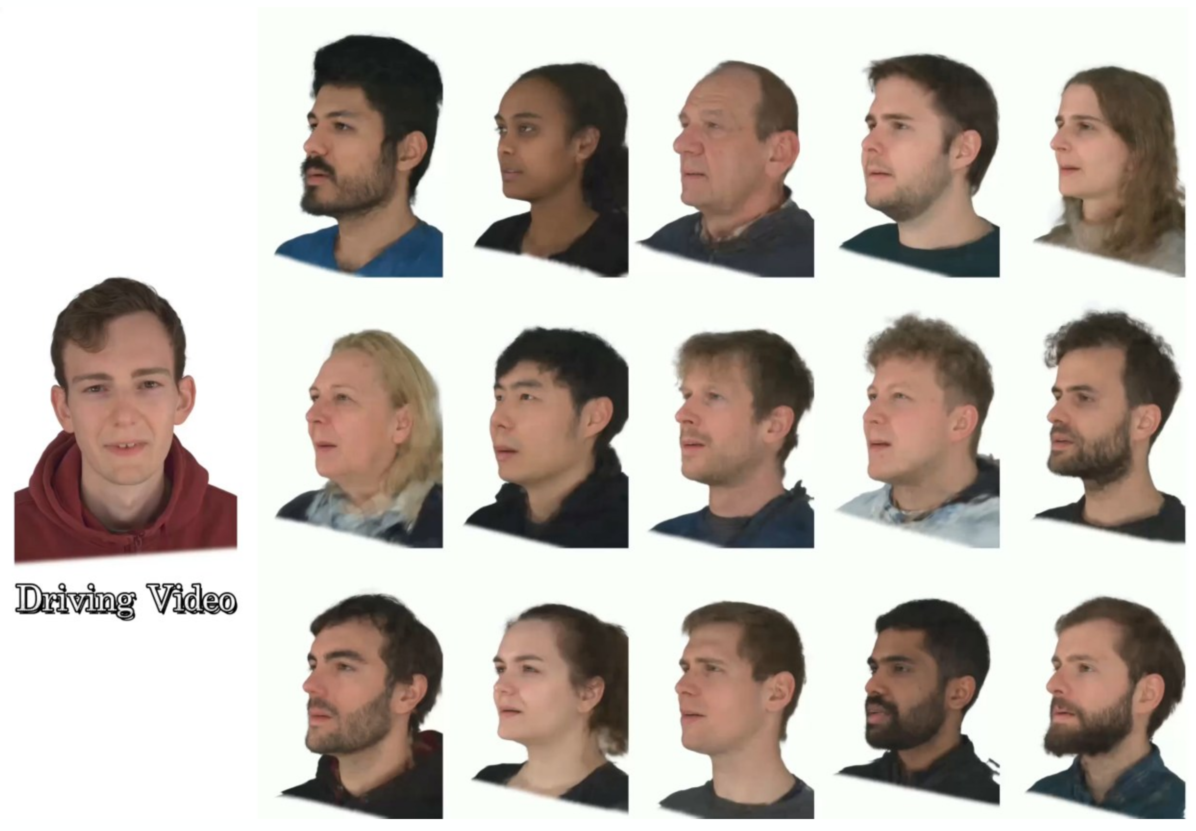
3.1 Animatable Avatar Creation

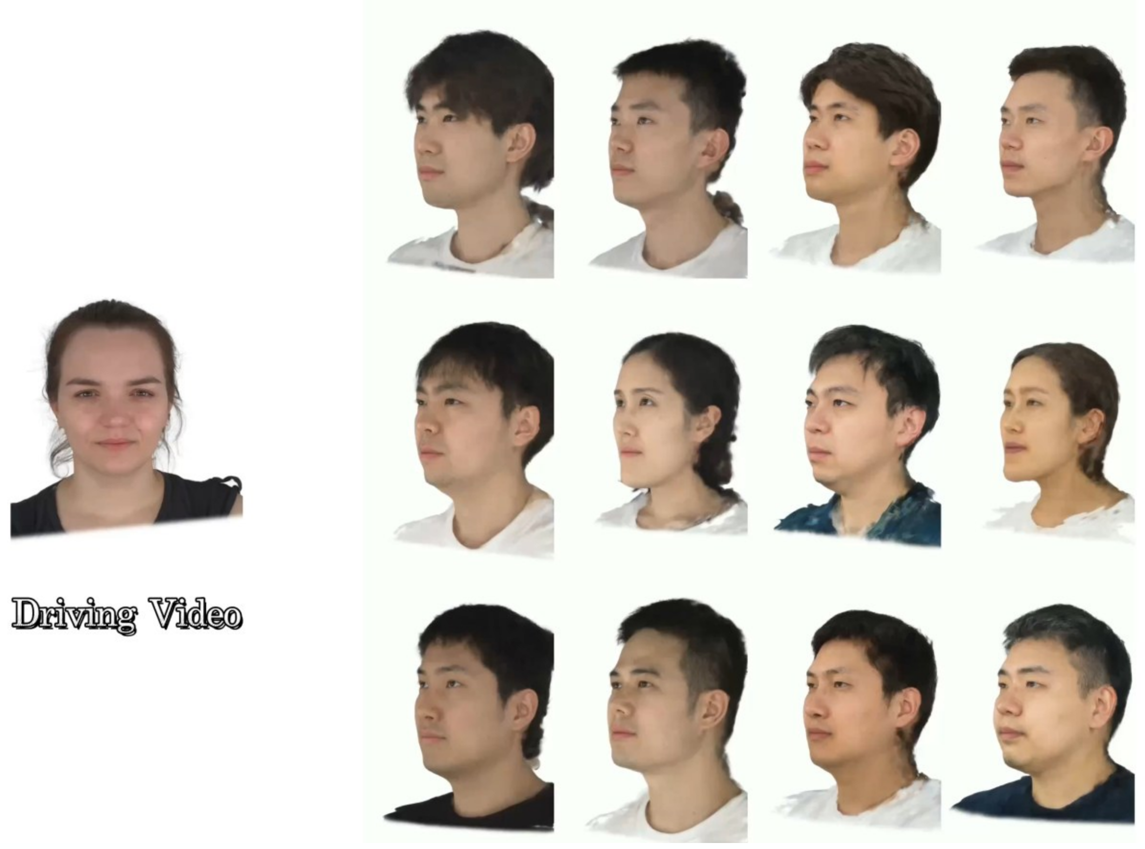
b. HeadGAP: Few-shot 3D Head Avatar via Generalizable Gaussian Priors Instant (3DV’25)

Personalization | Inversion and Finetuning

Achieved fast head avatar personalization by leveraging inversion and fine-tuning strategies using only few-shot images.







Prior Model | Auto-decoder framework

Presented auto-decoder designs that effectively utilize part-based dynamic Gaussian head priors trained from multi-ID, multi-view and multi-expression data.

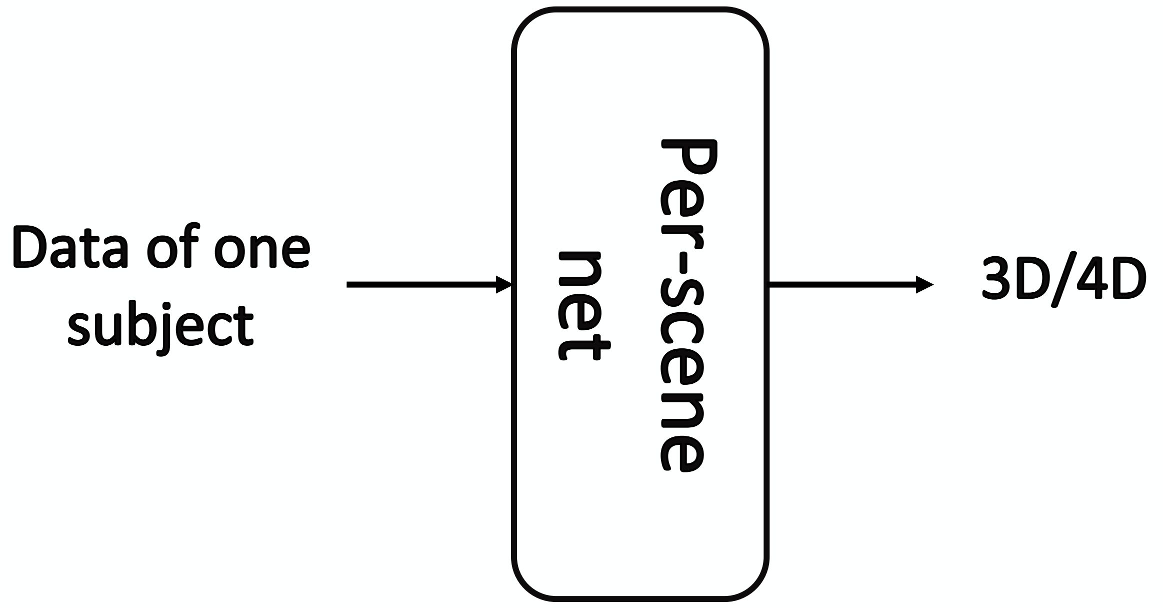
First to leverage generalizable 3DGS priors for few-shot head avatar creation

3. 形象重建：从先验模型到3D生成

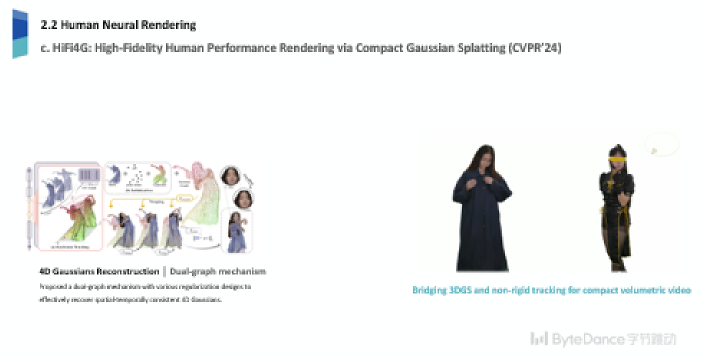
3. Avatar Reconstruction：From Prior model to 3D Generation

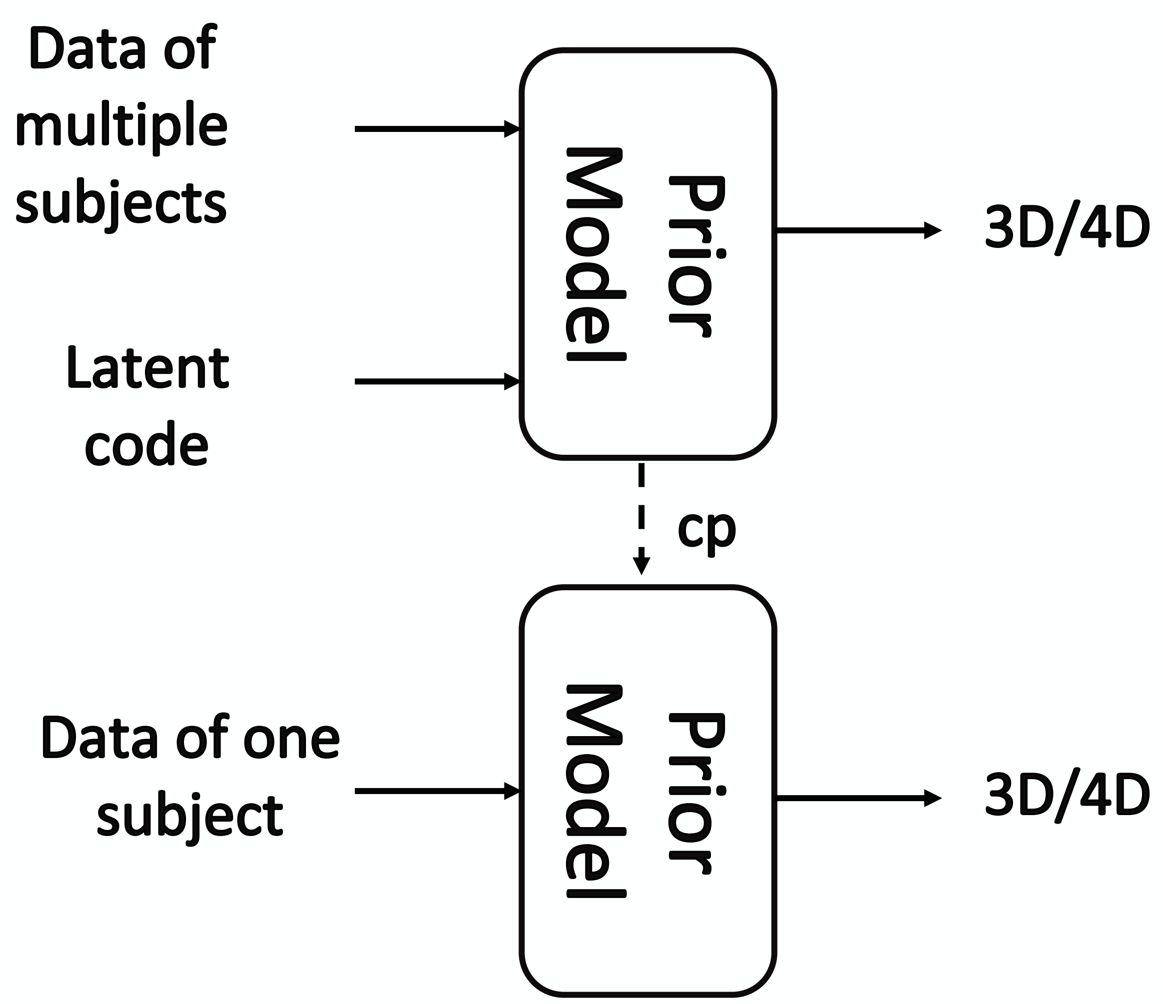
3.2 基于生成模型的三维重建

3.2 Human 3D Generation

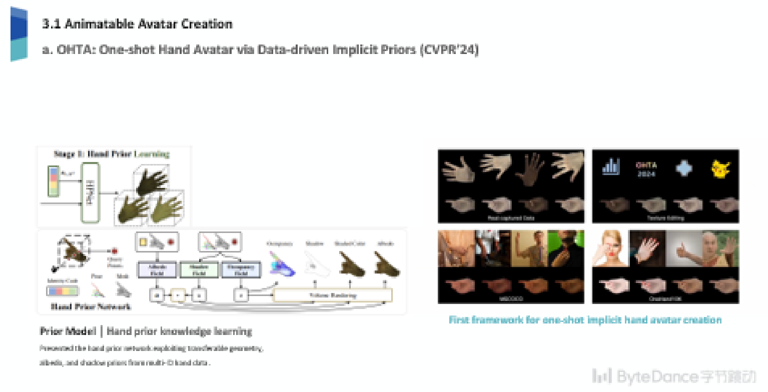


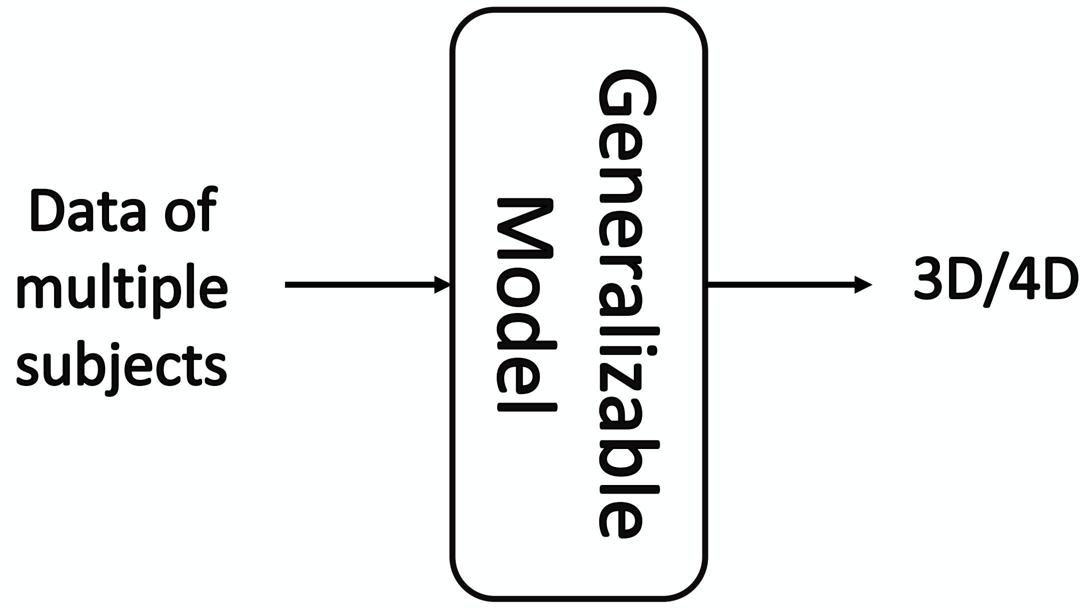
Per-scene training



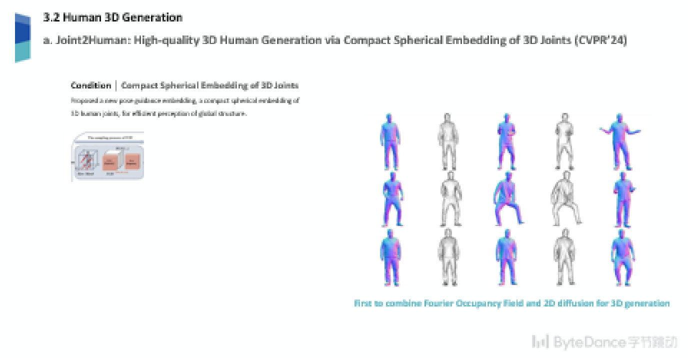


Prior training & tuning





Generalizable training

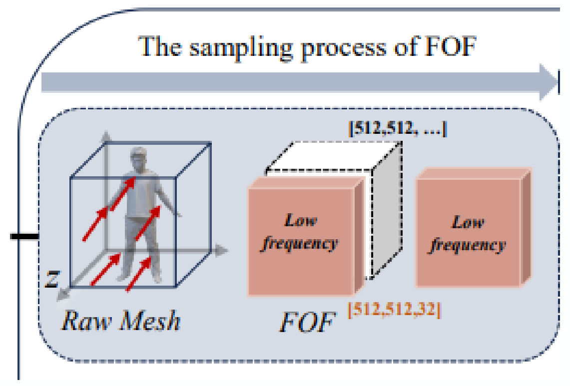


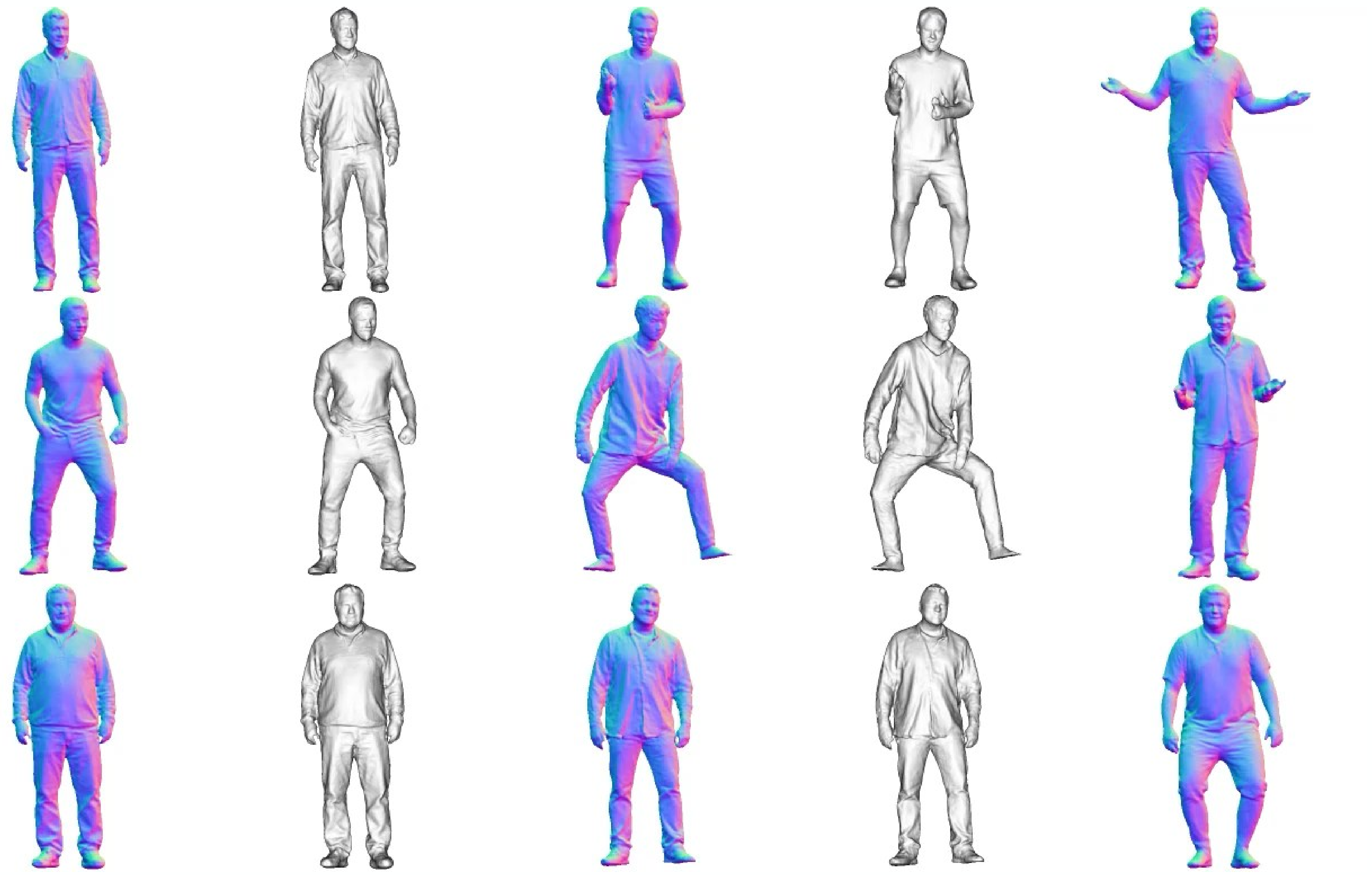
3.2 Human 3D Generation

a. Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints (CVPR’24)

Condition | **Compact Spherical Embedding of 3D Joints**

Proposed a new pose guidance embedding, a compact spherical embedding of 3D human joints, for efficient perception of global structure.





First to combine Fourier Occupancy Field and 2D diffusion for 3D generation

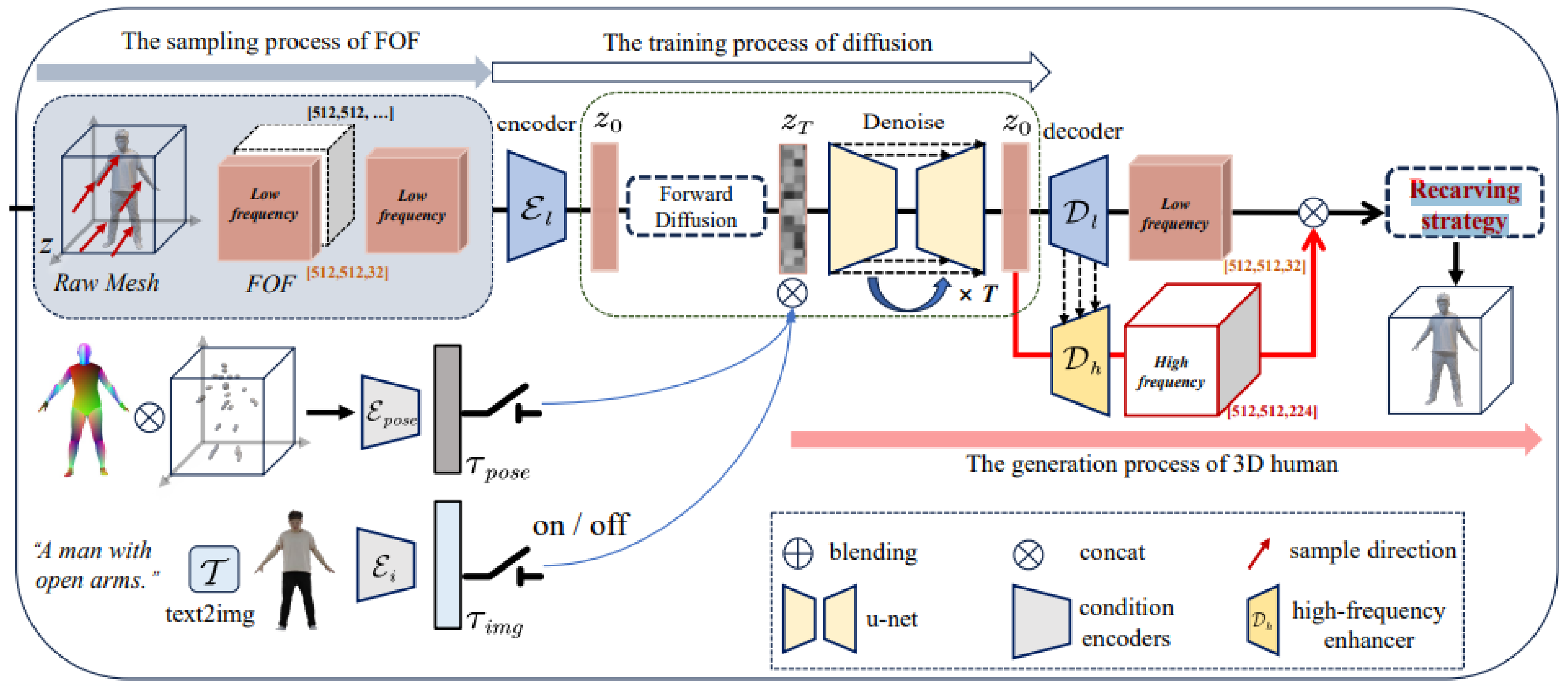
ByteDance字节跳动

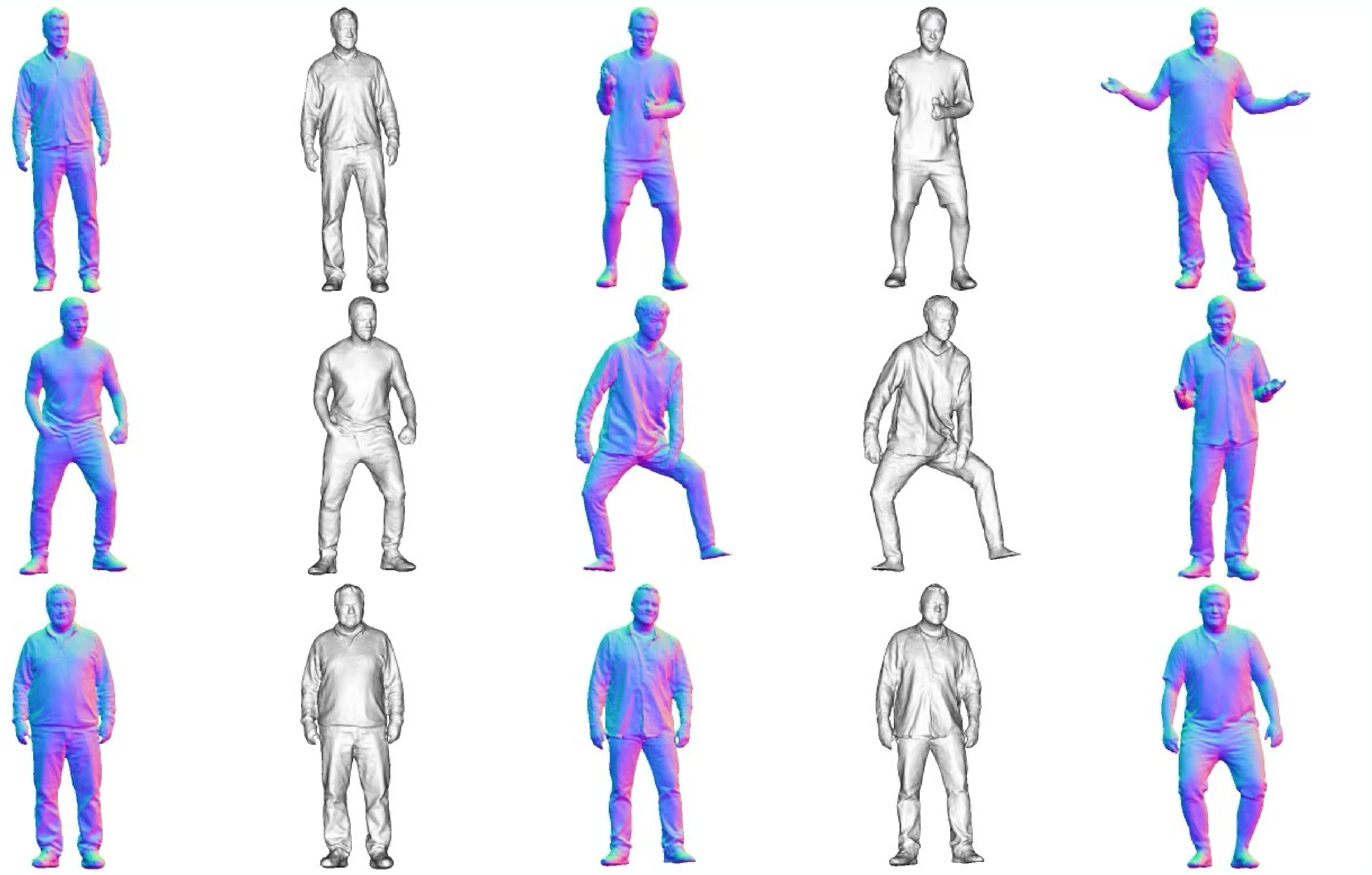
3.2 Human 3D Generation

a. Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints (CVPR'24)

Condition | Compact Spherical Embedding of 3D Joints

Proposed a new pose guidance embedding, a compact spherical embedding of 3D human joints, for efficient perception of global structure.





First to combine Fourier Occupancy Field and 2D diffusion for 3D generation

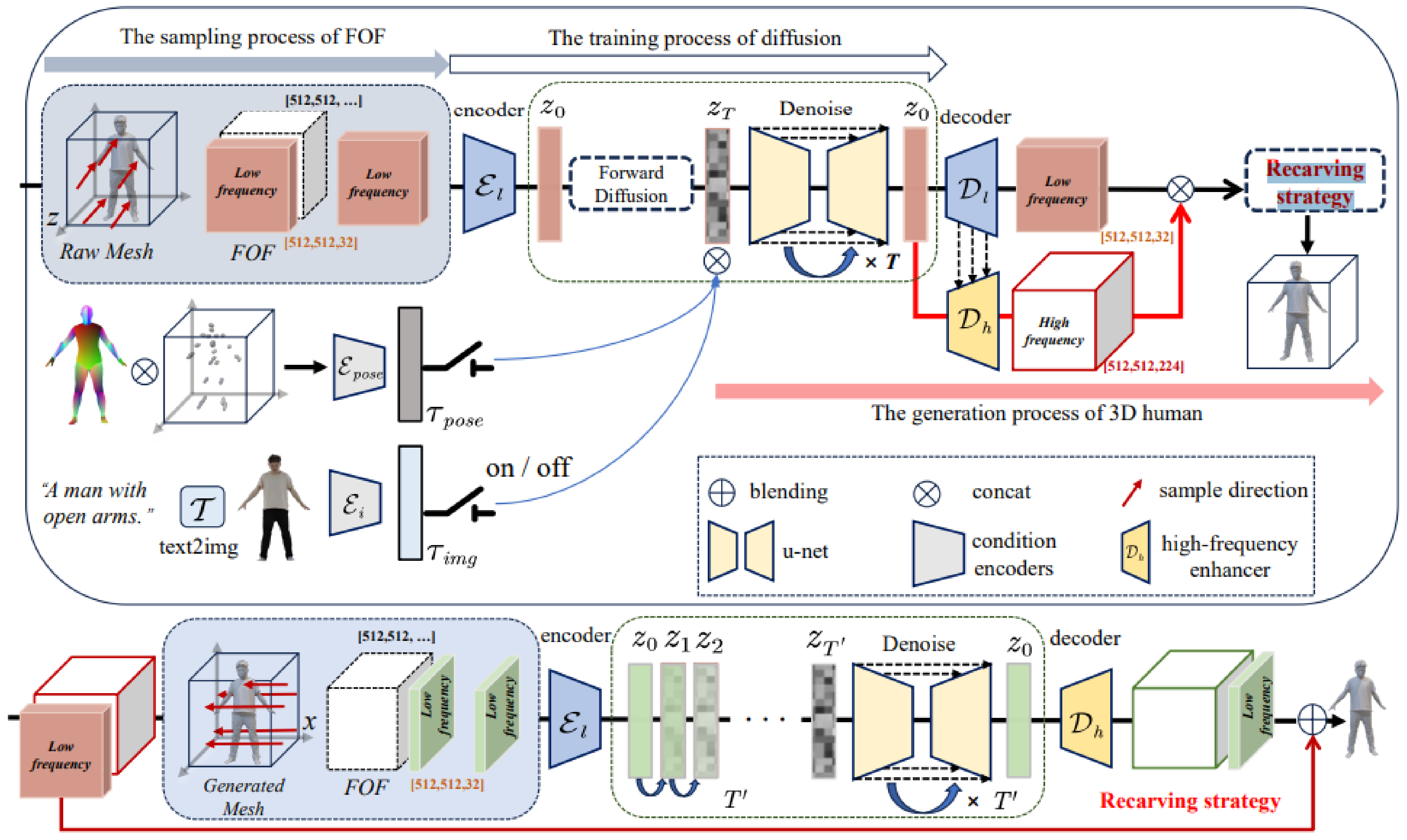
ByteDance字节跳动

3.2 Human 3D Generation

a. Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints (CVPR’24)

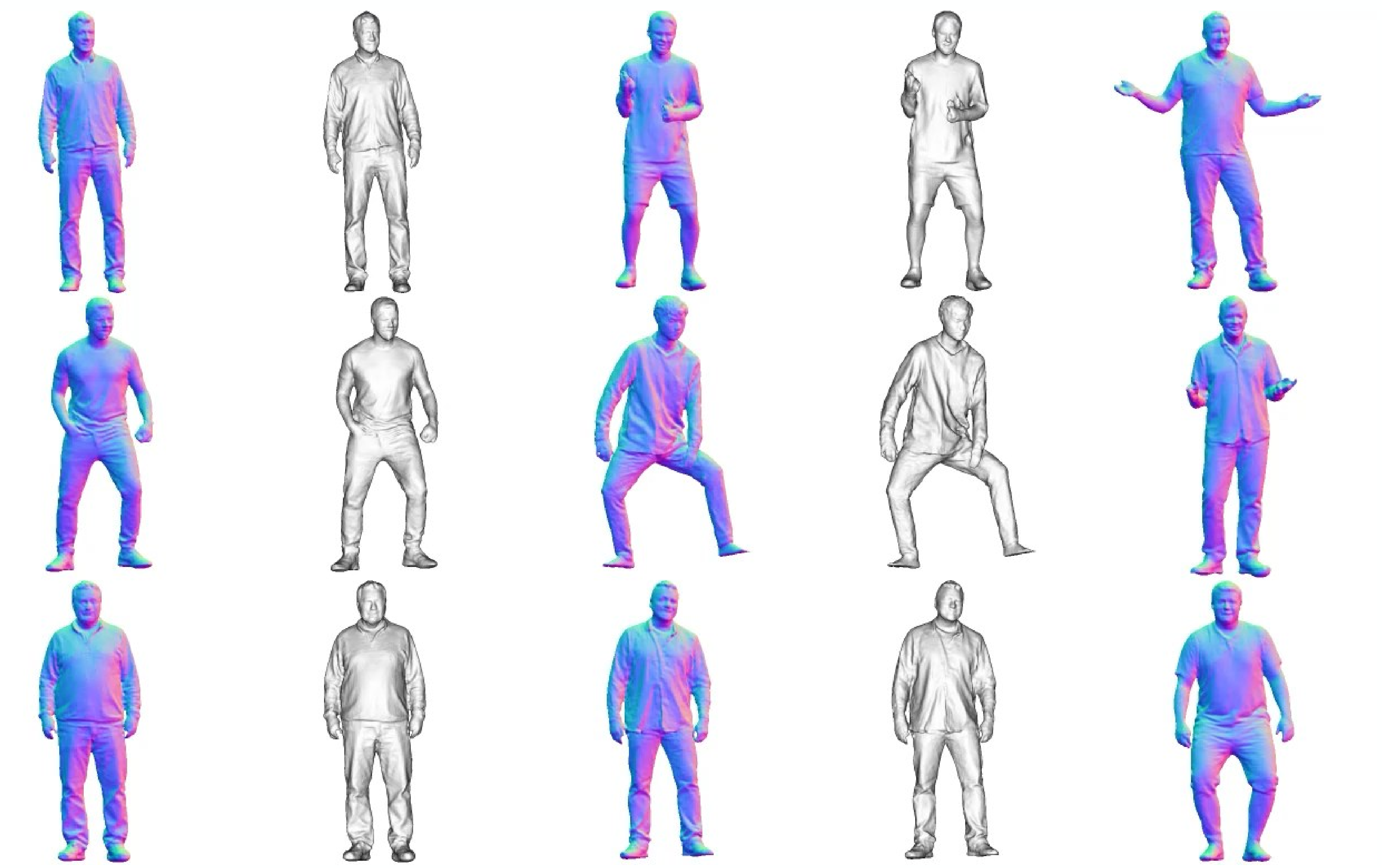
Condition | Compact Spherical Embedding of 3D Joints

Proposed a new pose guidance embedding, a compact spherical embedding of 3D human joints, for efficient perception of global structure.



Fine-grained Generation | Multiview recurving strategy

Designed a high-frequency enhancer by integrating a subsidiary decoder into the pre-trained VAE and a multiview recarving strategy for fine-grained local detail generation.



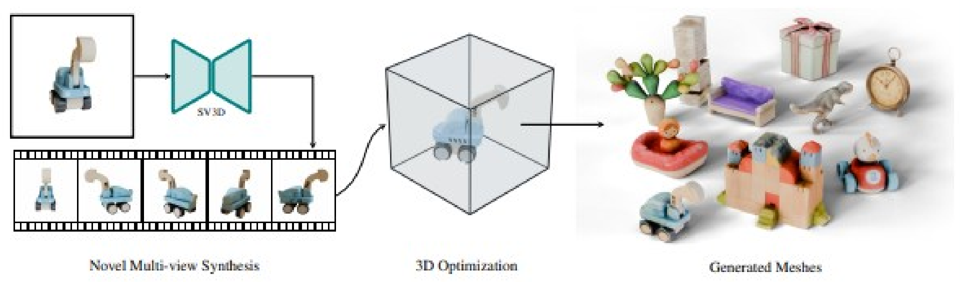
First to combine Fourier Occupancy Field and 2D diffusion for 3D generation

3.2 Human 3D Generation

b. HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors (NeuIPS’24)

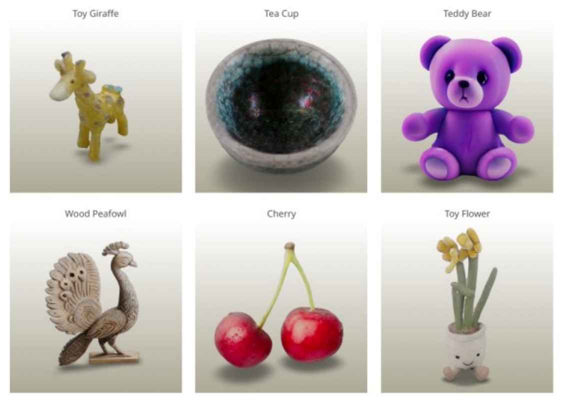
b-1 Background

LU & Microsoft SORA

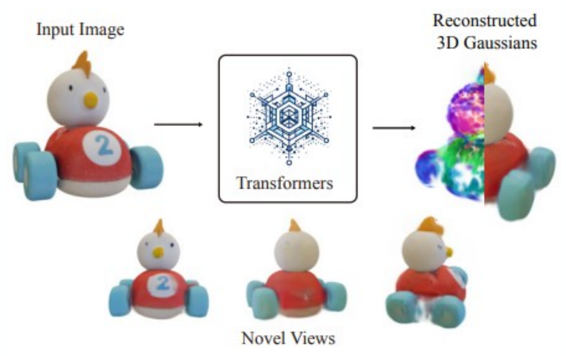


Stability AI SV3D

Video Generation

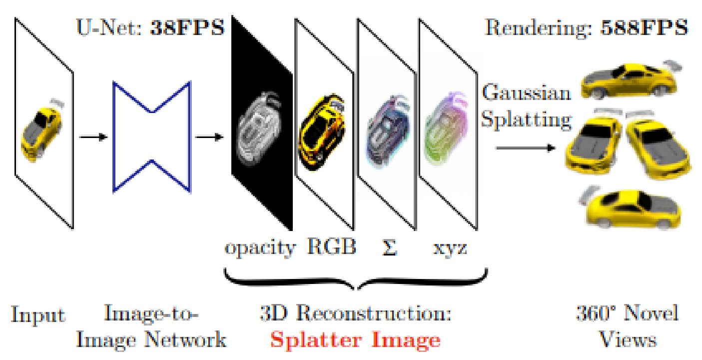


Adobe LRM

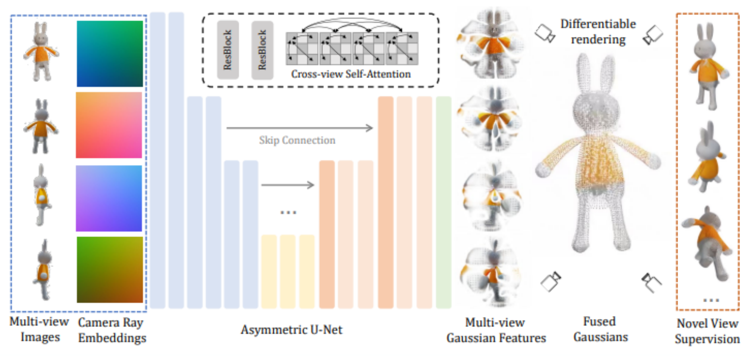


THU Triplane-Meets-GS

Large Reconstruction model



Oxford Splatter-Image



PKU LGM

Image-based 3DGS

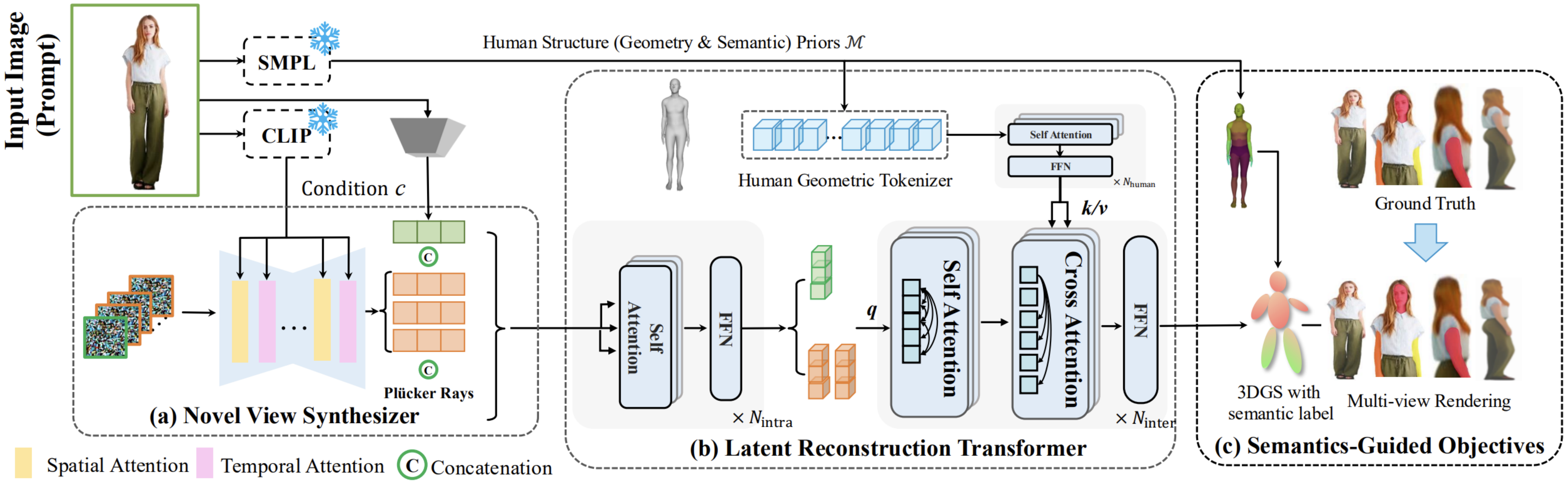
ByteDance 字节跳动

3.2 Human 3D Generation

b. HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors (NeuIPS’24)

b-2 Framework

First to leverage latent Gaussian reconstruction with a 2D generative diffusion model and 3D structure priors for efficient, high-fidelity single-image human reconstruction in an end-to-end framework.



Video Generation + Large Reconstruction Model

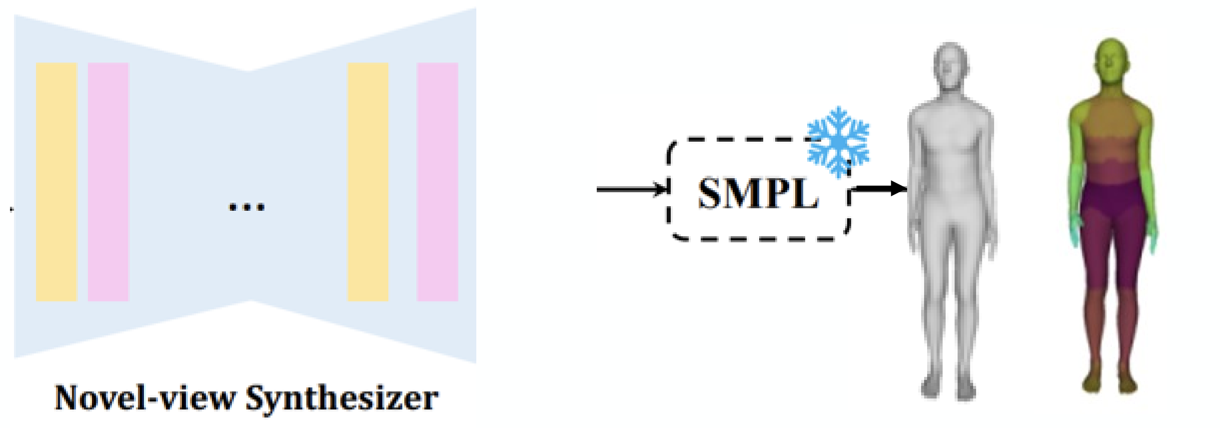
ByteDance字节跳动

3.2 Human 3D Generation

b. HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors (NeuIPS’24)

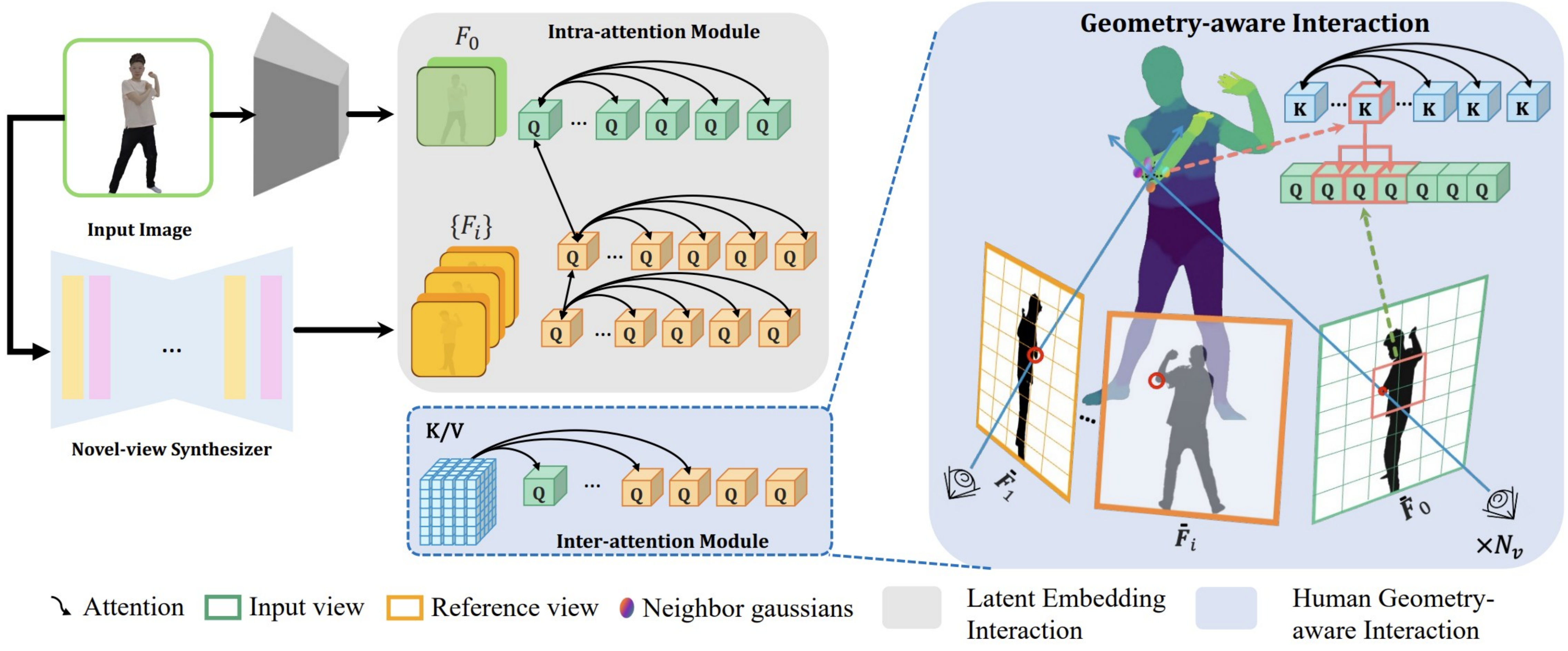
b-3 Core Design

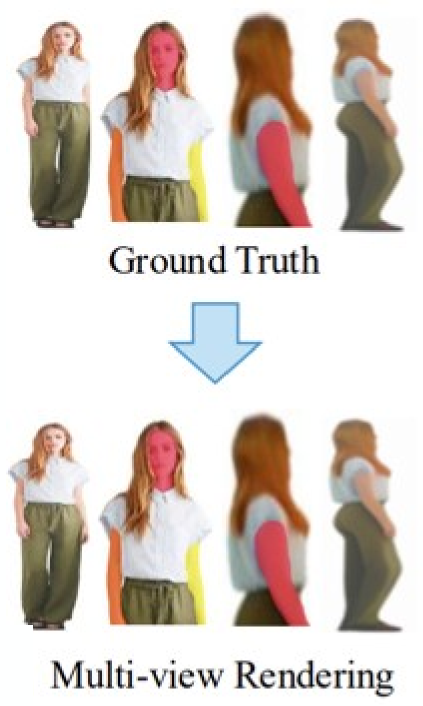
We integrate structure, semantic and appearance cues within a universal Transformer framework, leveraging SMPL geometry priors to stabilize high-quality human geometry generation and 2D generative diffusion appearance priors to hallucinate unseen parts of humans.



Appearance cues

Structure cues





Semantic-cues-guided loss

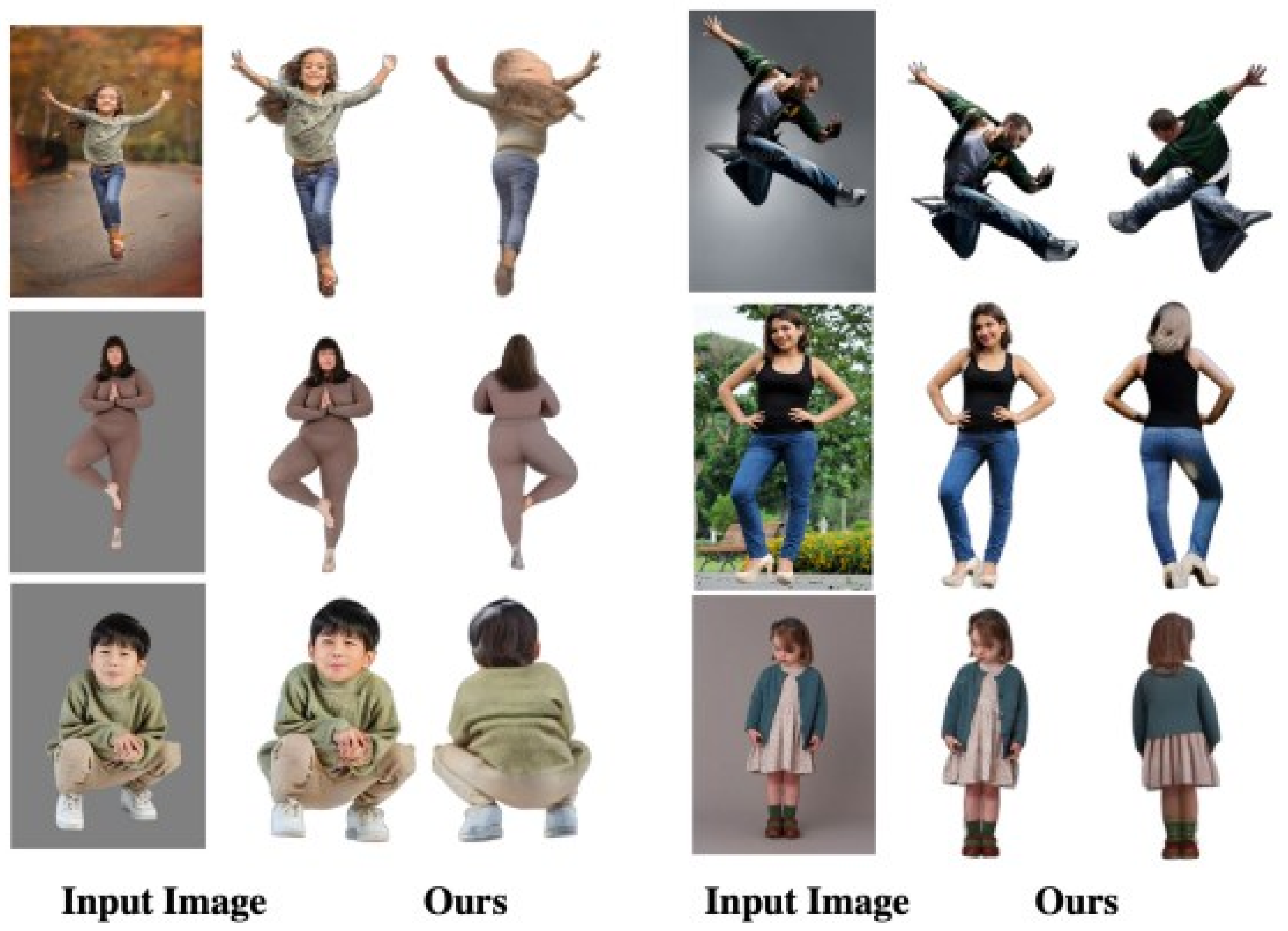
Latent Reconstruction Transformer

ByteDance字节跳动

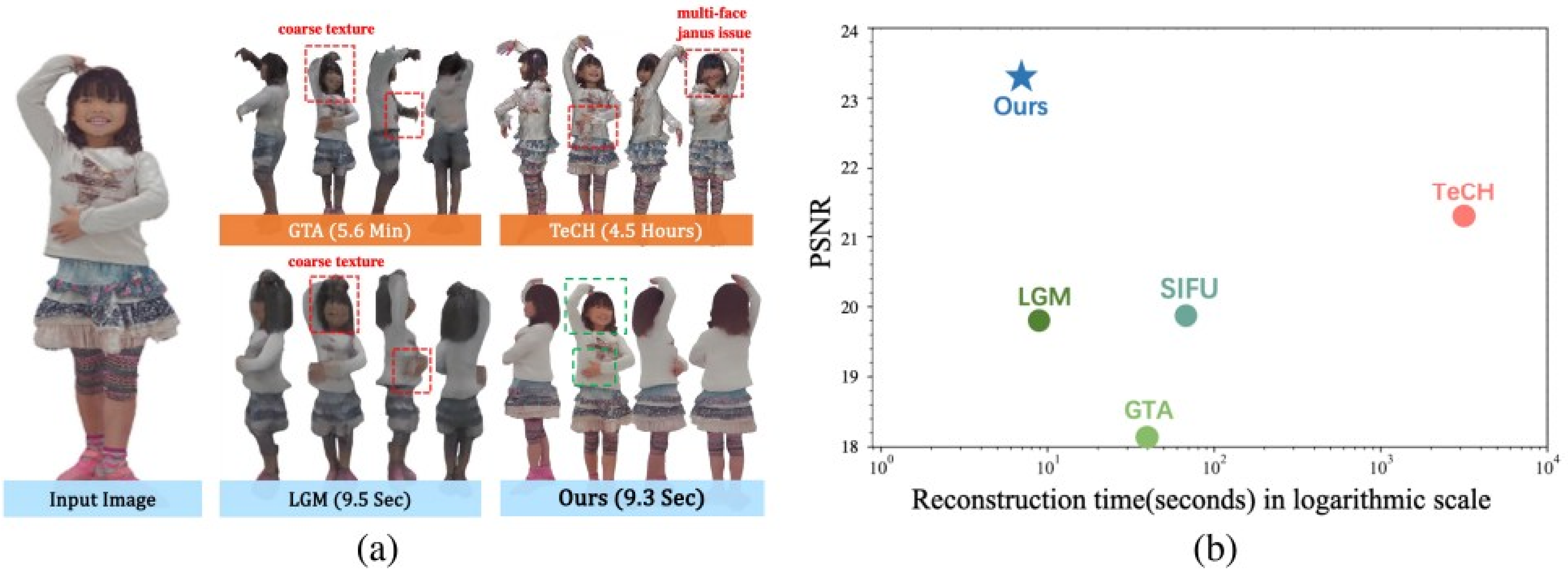
3.2 Human 3D Generation

b. HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors (NeuIPS’24)

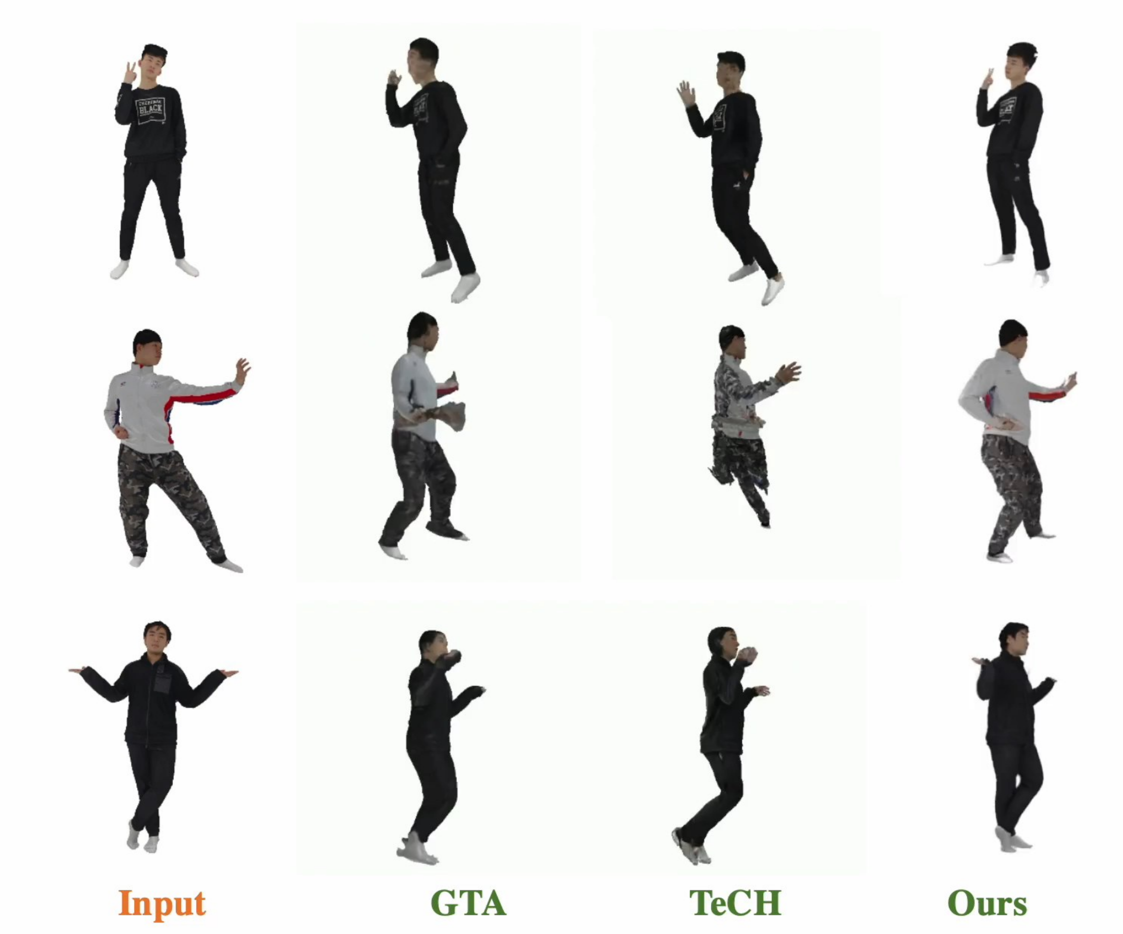
b-4 Results



In-the-wild Evaluation



Quantitative Evaluation



Video Results

3. 形象重建：从先验模型到3D生成

3. Avatar Reconstruction：From Prior model to 3D Generation

3.1 基于先验模型的形象重建

3.1 Animatable Avatar Creation

3.2 基于生成模型的三维重建

3.2 Human 3D Generation

Per-scene Reconstruction

Prior Model Training + Finetuning

Feedforward Generation

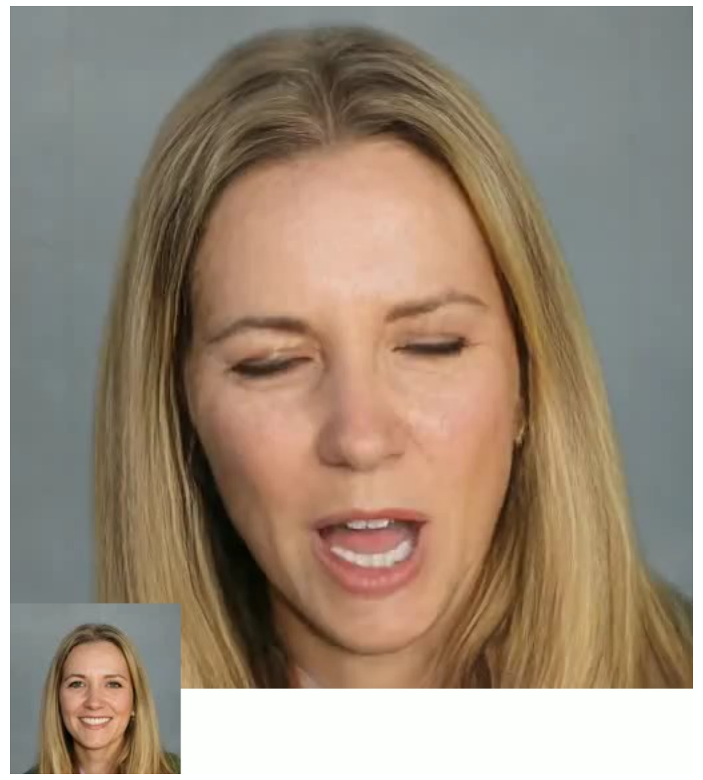
生成化

高效化

ByteDance字节跳动

Future work?

传统方法	神经渲染	先验模型	三维生成	化身生成	以人为中心的4D生成
2018.07 – 2021.06	2021.07 – now	2023.05 – now	2023.06 – now	2024.12-future	2024.12-future
4D Volumetric Capture	4D Neural Rendering	Avatar creation via Prior model	Human 3D Generation	Realistic 3D Avatar Generation	Human-centric 4D Generation



Microsoft VASA1

THU & MIT Genesis

World Labs Generating Worlds

ByteDance字节跳动

Appendix

Reported paper with link

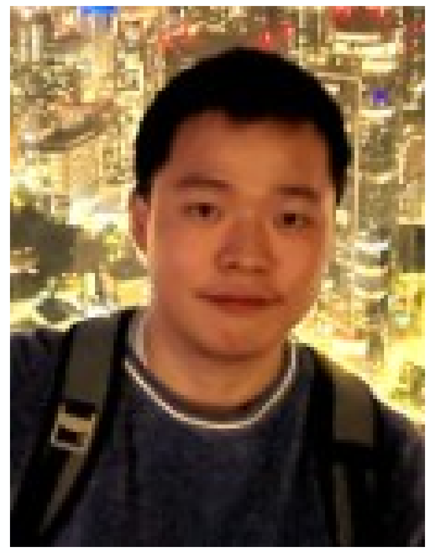
1. UnstructuredFusion: Realtime 4D Geometry and Texture Reconstruction using Commercial RGBD Cameras
2. RobustFusion: Human Volumetric Capture with Data-driven Visual Cues using a RGBD Camera
3. Robust Volumetric Performance Reconstruction under Human-object Interactions from Monocular RGBD Stream
4. NeuralHOFusion: Neural Volumetric Rendering under Human-object Interactions
5. Instant-NVR: Instant Neural Volumetric Rendering for Human-object Interactions from Monocular RGBD Stream Instant Neural Human and Object Rendering from Single RGBD Sensor
6. HiFi4G: High-Fidelity Human Performance Rendering via Compact Gaussian Splatting Instant Neural Human and Object Rendering from Single RGBD Sensor
7. OHTA: One-shot Hand Avatar via Data-driven Implicit Priors
8. HeadGAP: Few-shot 3D Head Avatar via Generalizable GAussian Priors
9. Joint2Human: High-quality 3D Human Generation via Compact Spherical Embedding of 3D Joints
10. HumanSplat: Generalizable Single-Image Human Gaussian Splatting with Structure Priors

Acknowledgements


Lu Fang


Yebin Liu


Lan Xu


Yuheng Jiang


Xiaozheng Zheng


Kun Li


Muxin Zhang


Panwang Pan

Q&A

ByteDance字节跳动

THANKS

ByteDance字节跳动