OneGT: One-Shot Geometry-Texture Neural Rendering for Head Avatars

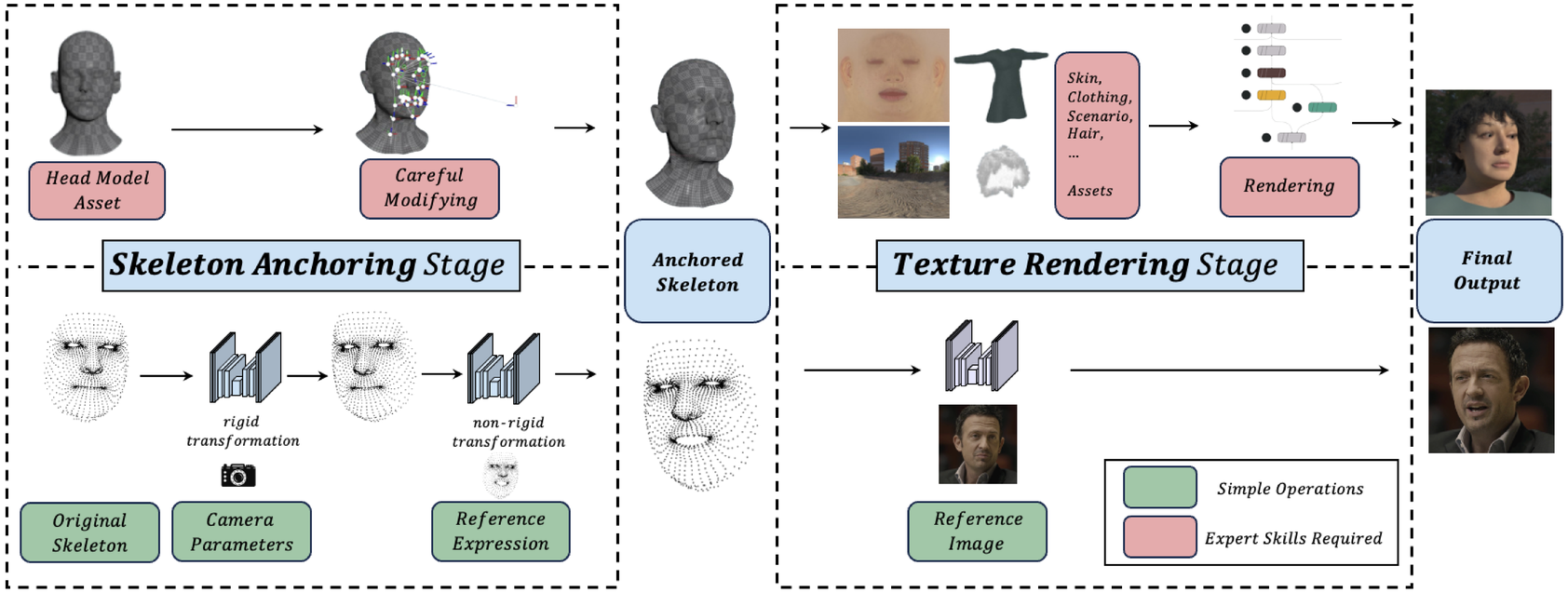
Jinshu Chen, Bingchuan Li†, Fan Zhang, Songtao Zhao, Qian He

Intelligent Creation Team, ByteDance

{chenjinshu, libingchuan, noodles, zhaosongtao.0815, heqian}@bytedance.com

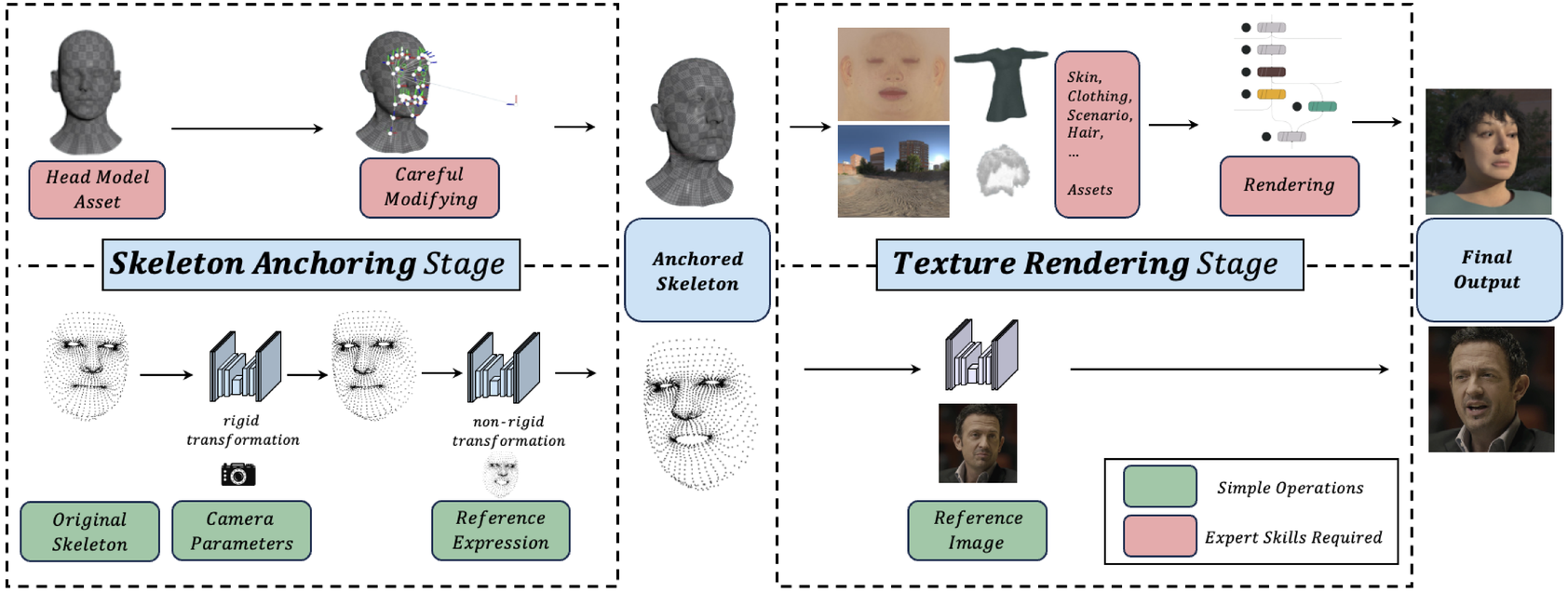
Summary

- Existing solutions for creating high-fidelity digital head avatars encounter various obstacles.
  - Traditional rendering tools offer realistic results, while heavily requiring expert skills; Neural rendering methods are more efficient but often compromise between the generated fidelity and flexibility.
- We present ***OneGT*** that adheres to the frameworks of the rendering tools, while restructuring individual stages of the rendering pipeline through neural networks.
  - Specifically, ***OneGT*** contains a skeleton-anchoring stage and a texture-rendering stage, in which well-designed Transformers learn the geometric transformations and the proposed reference-perceptible DiT renders the textures respectively.
  - Our framework learns geometric consistency from the innovatively introduced synthetic data, thus achieving superior performance while requiring only 10%-30% of the real-world data typically used by competitive methods.
  - Experimental results demonstrate that ***OneGT*** achieves high fidelity in producing portrait avatars, meanwhile maintaining the flexibility of editing.



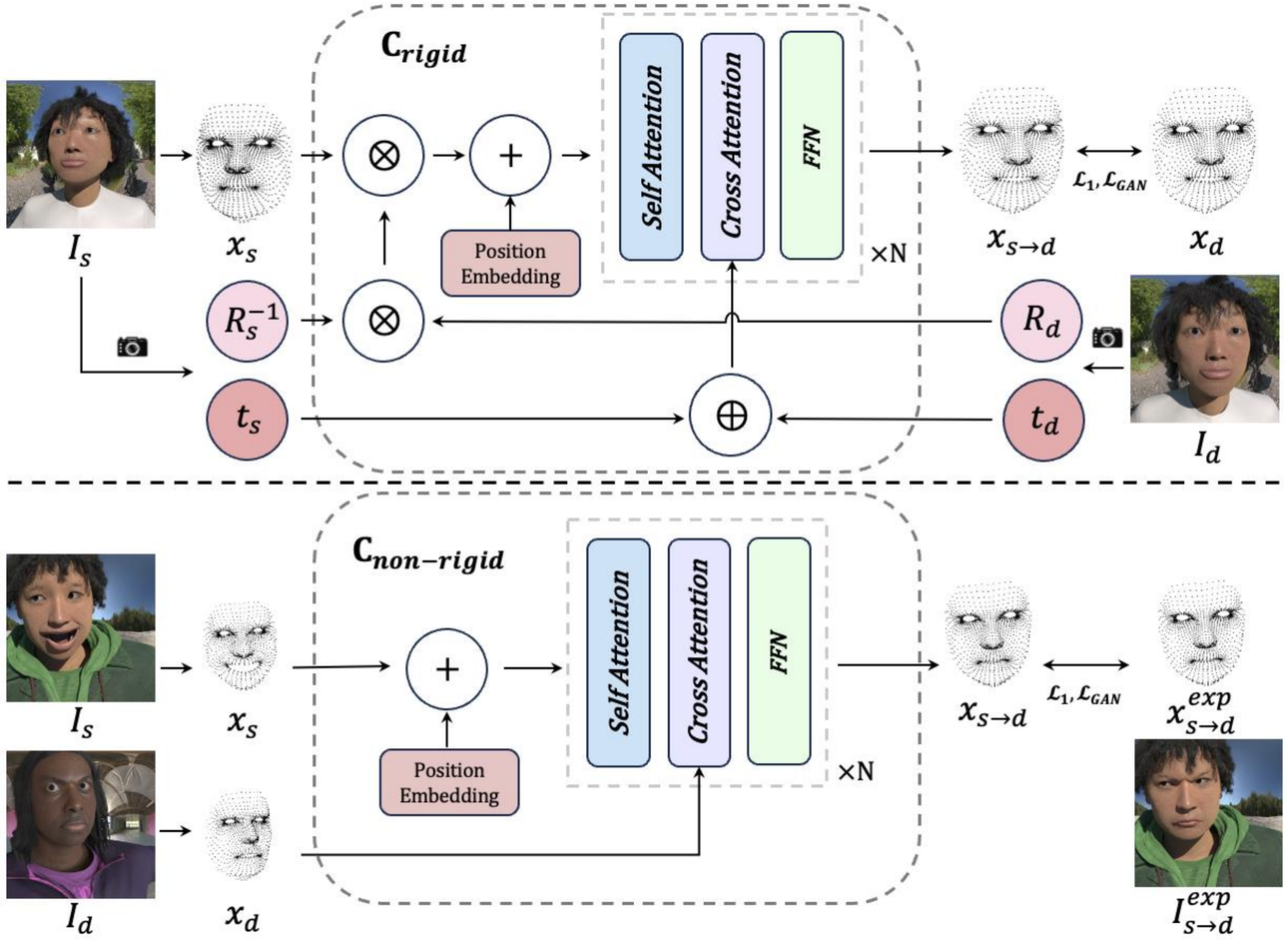
Highlights

- OneGT, a one-shot neural rendering framework architected according to the pipeline of the traditional rendering software.
- A geometry-texture decoupled pipeline: (1) the prior skeleton-anchoring stage contains two well-designed Transformers to comprehend the geometry; (2) the posterior texture-rendering stage is built upon the proposed reference-perceptible DiT.
- Instead of learning from massive real-world data, OneGT manages to gain geometric consistency from the innovatively introduced synthetic data, mitigating the required amount of the real-world datasets to 10%-30%.



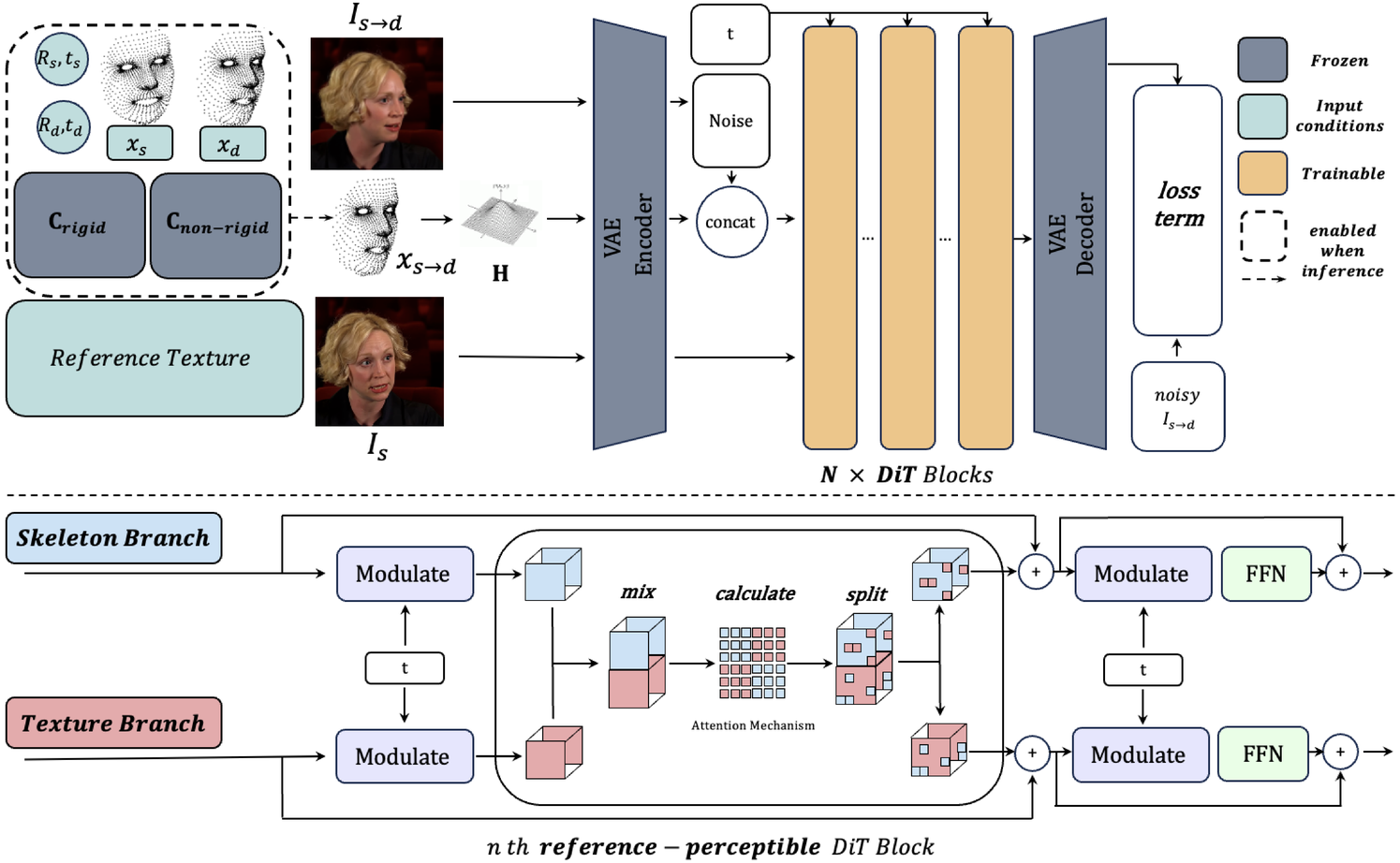
Framework

- OneGT contains two distinct phases: a preliminary skeleton-anchoring stage and a subsequent texture-rendering stage.
  - The former stage, skeleton anchoring, handles the generation of precise skeletal structures based on provided control conditions, e.g., camera parameters.
  - For the skeleton anchoring, we adopt the dense representation along with our introduced adaptive strategies as the skeleton, which reaches a precise balance between the fidelity and flexibility.
  - The skeleton is anchored through two specially designed Transformers which manage the rigid and non-rigid transformations respectively.



Framework

- OneGT contains two distinct phases: a preliminary skeleton-anchoring stage and a subsequent texture-rendering stage.
  - For the texture rendering, we propose our reference-perceptible DiT backbone, which contains a main branch taking in the skeleton information and a parallel branch handling the provided reference texture.
  - The texture information and the structural geometry get blended deeply through the designed "mix-calculate-split" attention modules, ensuring refined final results.



Dataset

- We incorporate synthetic data into our pipeline, which has high fidelity, cost-effective acquisition, and high customizability.
  - Unlike existing frameworks that attempt to simultaneously learn geometric structures and texture details from massive real-world datasets, our architecture moves the part of geometric consistency entirely to synthetic data.
  - This approach enables our framework to achieve superior performance while requiring only 10%-30% of the real-world data consumed by the competitive peers.

Rendering Data

- We use Houdini to manufacture the rendering data.
- Specifically, over 200 different preset head models, skin textures, accessories and background assets are stored and randomly combined.
  - we produce about 50,000 groups of the data, each of which contains 20 images with the same identity, posture, expression, scene, but different camera perspectives.
  - C_{rigid} is trained completely on this mentioned data.
  - Such rendering data also partially supports the training of the texture rendering stage.

Dataset

- We incorporate synthetic data into our pipeline, which has high fidelity, cost-effective acquisition, and high customizability.
  - Unlike existing frameworks that attempt to simultaneously learn geometric structures and texture details from massive real-world datasets, our architecture moves the part of geometric consistency entirely to synthetic data.
  - This approach enables our framework to achieve superior performance while requiring only 10%-30% of the real-world data consumed by the competitive peers.

Rendering Data

- We use Houdini to manufacture the rendering data.
- Specifically, over 200 different preset head models, skin textures, accessories and background assets are stored and randomly combined.
  - we render 0.5 million triplet groups, applying the translation coefficients of I_d directly to I_s to obtain $I_{s\to d}^{exp}$. Note that I_s differs from $I_{s\to d}^{exp}$ only in expressions, but rarely has similar attributes compared to I_d.
  - Such triplet data constructs the whole training set of $C_{non-rigid}$.

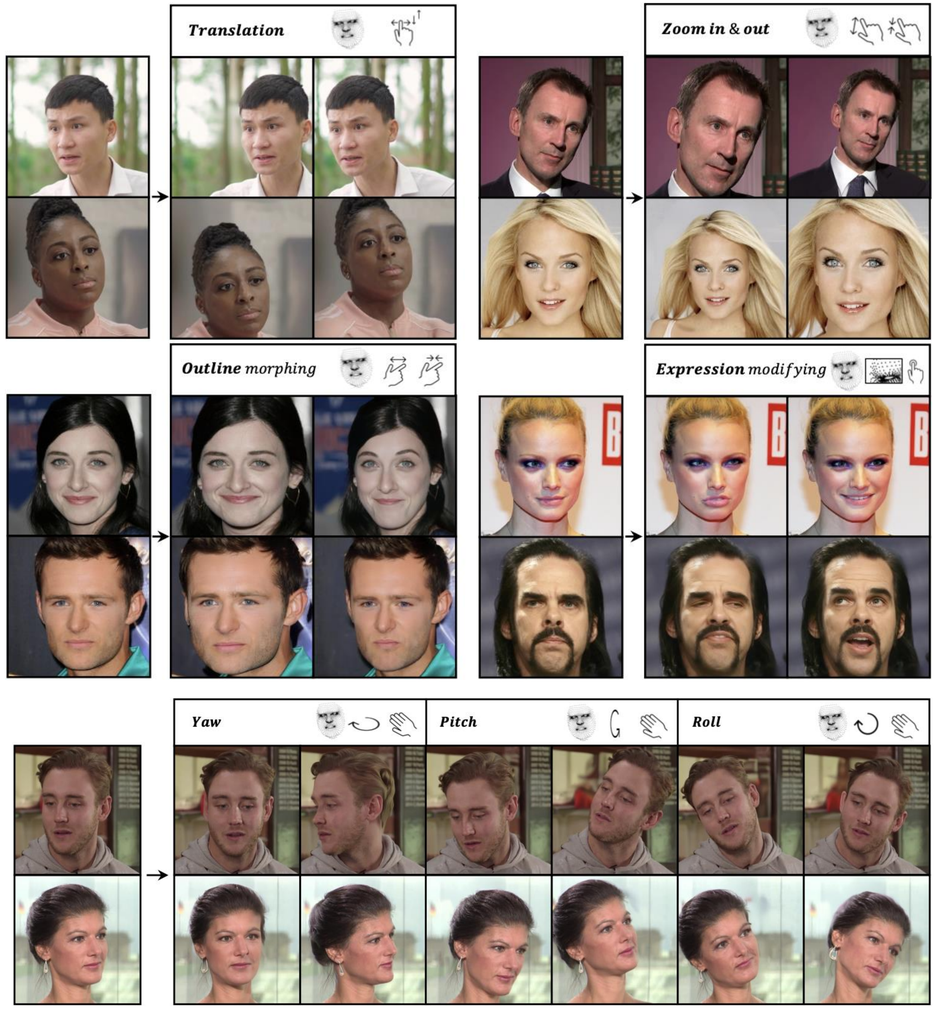
Dataset

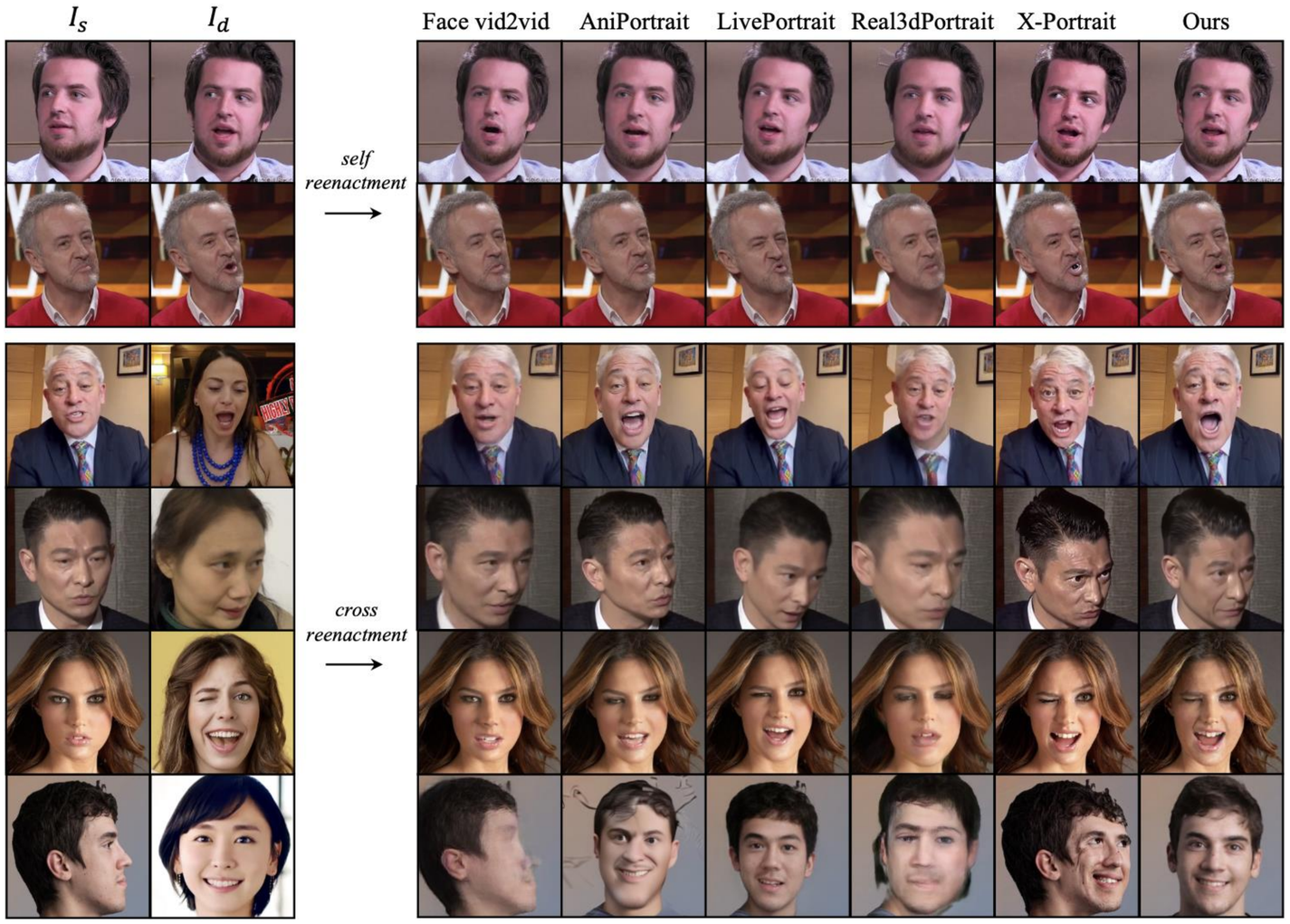
- We incorporate synthetic data into our pipeline, which has high fidelity, cost-effective acquisition, and high customizability.
  - Unlike existing frameworks that attempt to simultaneously learn geometric structures and texture details from massive real-world datasets, our architecture moves the part of geometric consistency entirely to synthetic data.
  - This approach enables our framework to achieve superior performance while requiring only 10%-30% of the real-world data consumed by the competitive peers.

Read-world Data

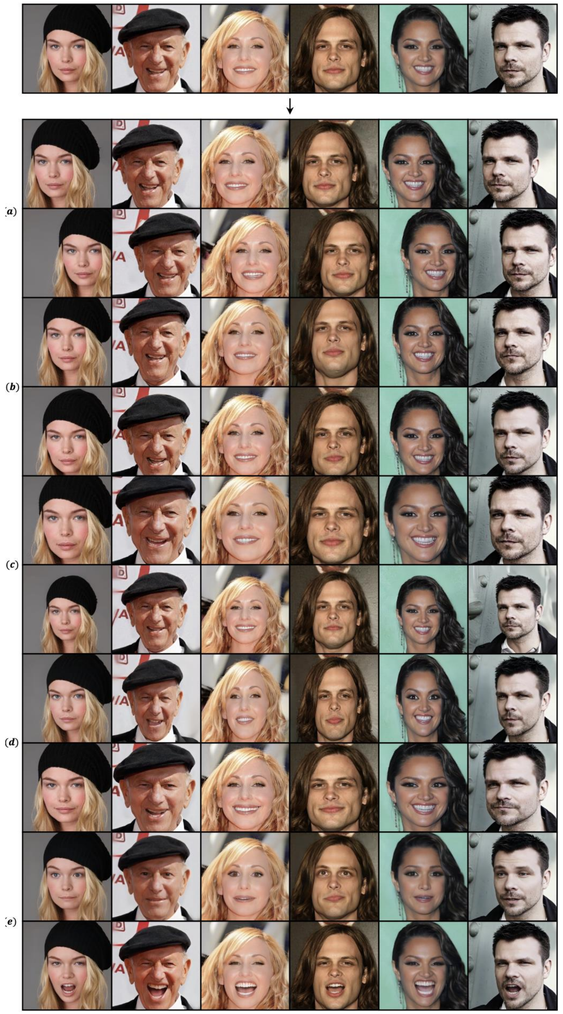
- We construct the real-world domain training data from VFHQ.
- Based on the raw data, we roughly screen about 1 million single-frame data according to the principles of no facial occlusion and static background..
- Generally, the real-world domain data only participates in the training of texture rendering, leaving the accountability for learning the geometry to the rendering data..

Results

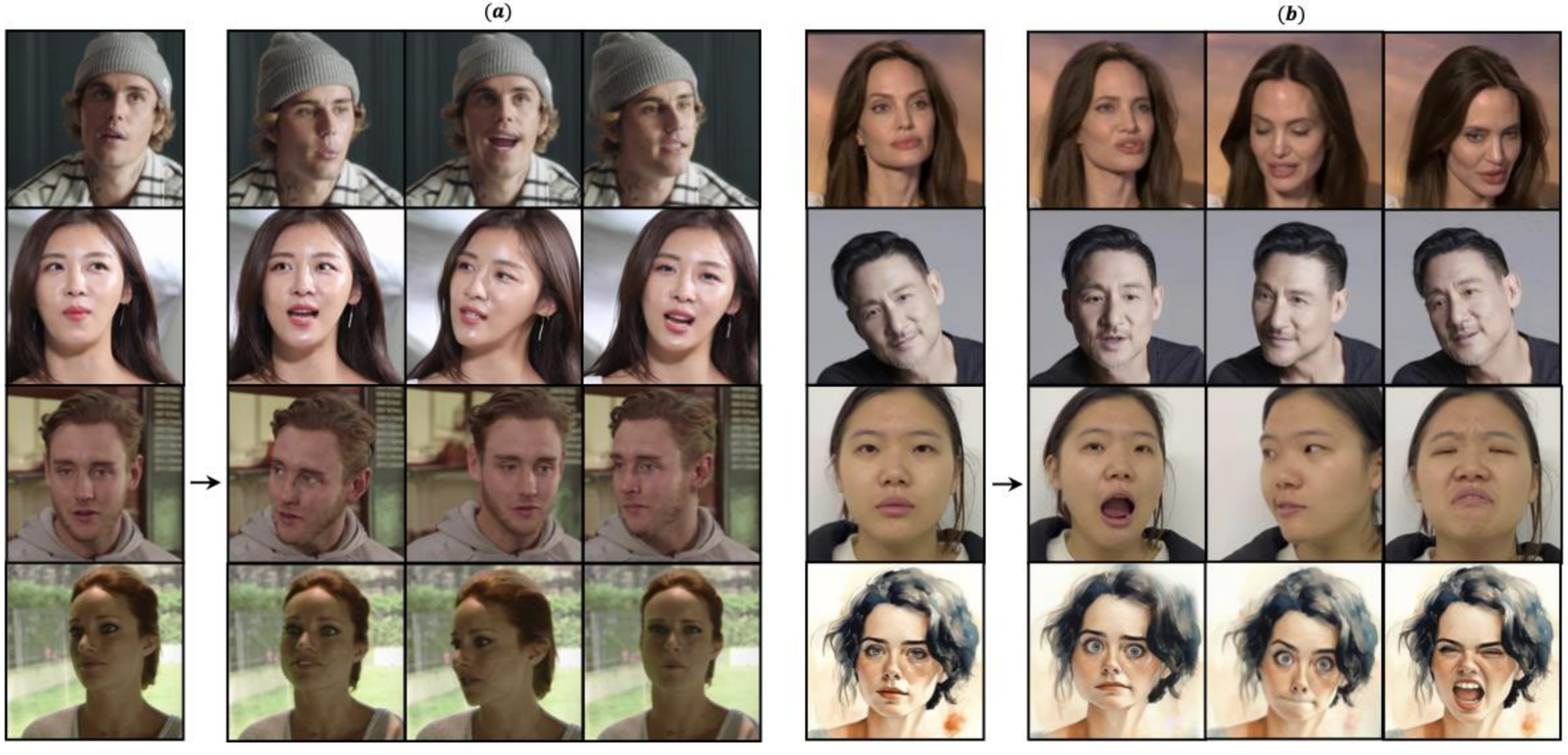

Translation
Zoom in & out
Outline morphing
Expression modifying
Yaw
Pitch
Roll


I_s
I_d
Face vid2vid
AniPortrait
LivePortrait
Real3dPortrait
X-Portrait
Ours
self reenactment
cross reenactment

Results



Results



ByteDance 字节跳动

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Thanks for watching!