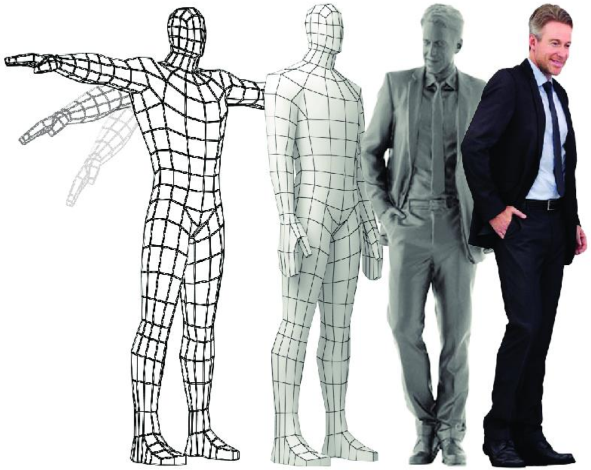
RECENT ADVANCES IN AVATARS

David Moreno, **Peter Eisert**, Fraunhofer HHI


Vision & Imaging Technologies

Innovative research along the entire video processing chain

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE

Imaging Technologies

Capturing Systems

Capturing & Processing

Computer Vision & Visual Computing

Image/ Video Analysis & Understanding

Modelling & Digital Representations

Generative Models & Animation

Visualization & Rendering

Imaging Technologies

Display & visualization technology

Reasoning

Interaction& HCI

Cognitive & Interactive Systems

Fields of Application

Industry & Construction

Medicine

Multimedia

Security

Avatars as Intuitive Interfaces

Applications

Virtual therapists

Employee- and communication training

Virtual teaching

Entertainment

Digital sign language specialist

Medicine & therapy

Virtual assistants

Realism vs Interactivity

Models

- ✚ Highly interactive, modifications possible
- ━ Realism is expensive (motion & appearance)

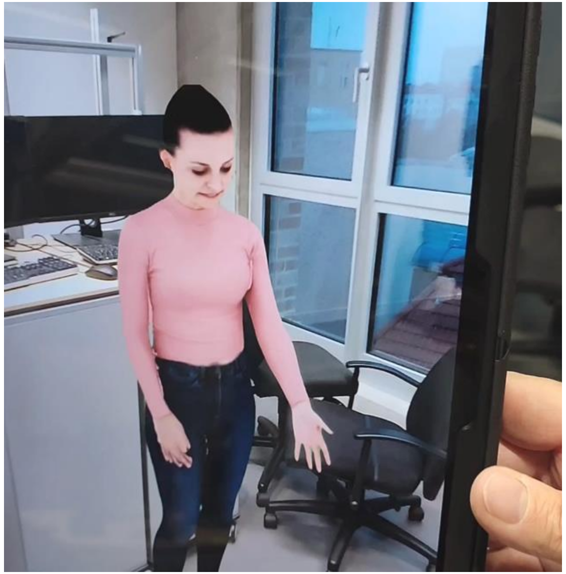
CG Data

Images, videos, volumetric video…

- ✚ Highly detailed and realistic
- ━ No modification possible

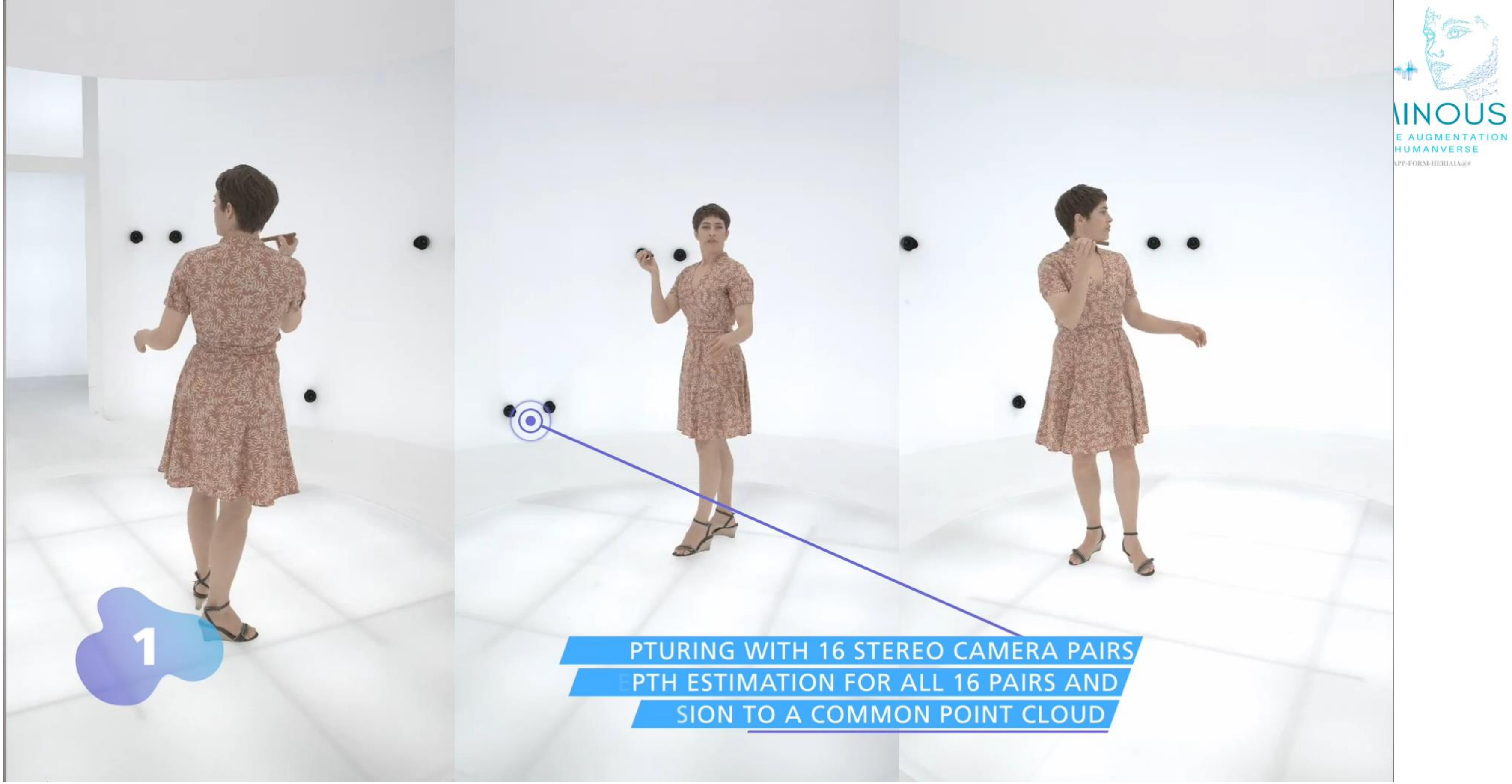
Real World Data

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE



- Creation of highly realistic representations of virtual humans that can naturally interact with users via text and speech
- Personalization of avatars to new target persons
- Style, semantic content and emotion adapted animation of virtual character

Volumetric Video Studio @ Fraunhofer HHI



A. Hilsmann et al, *IET Comp. Vision, 2020*

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE

Kinematic Animation of Volumetric Video

Head Rotation: VR Experience Eva Umlauf- Her Testimony

From Volumetric Video to Virtual Humans

Learning animatable representations from volumetric video

Volumetric Video with fitted model

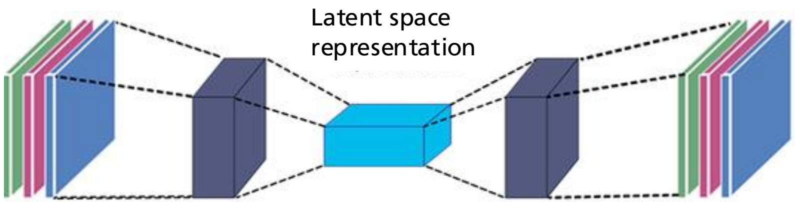
→

Novel pose,

Novel viewpoint,

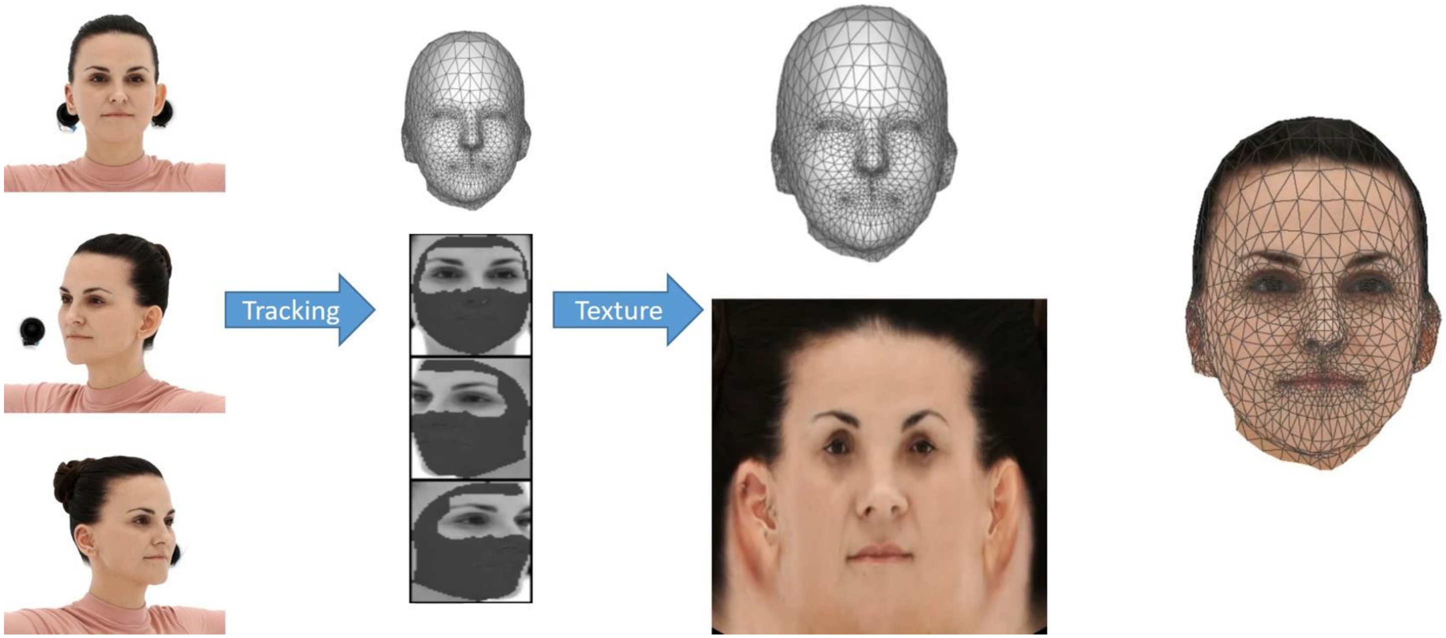
Novel facial expression

↓



Neural Face Representation

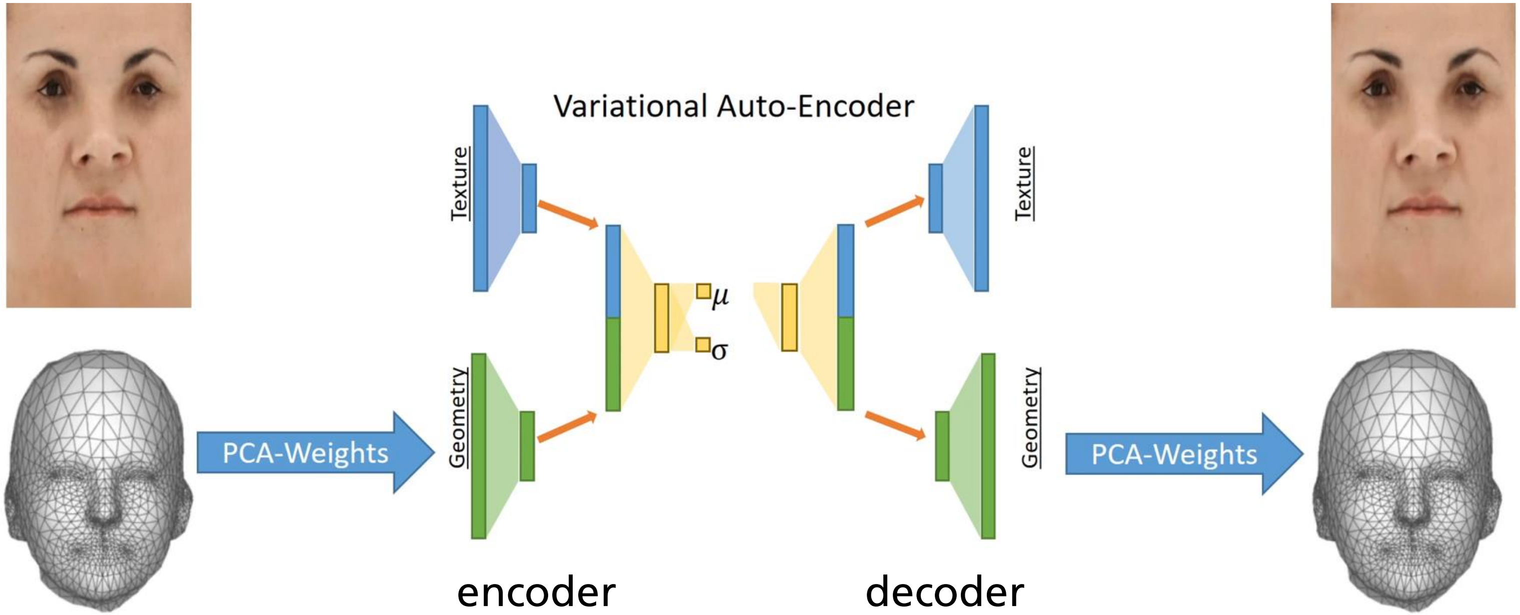
Hybrid face capture & representation learning



W. Paier et al., CVMP 2020
W. Paier et al., IET CV 2020

Neural Face Representation

Variational Autoencoder



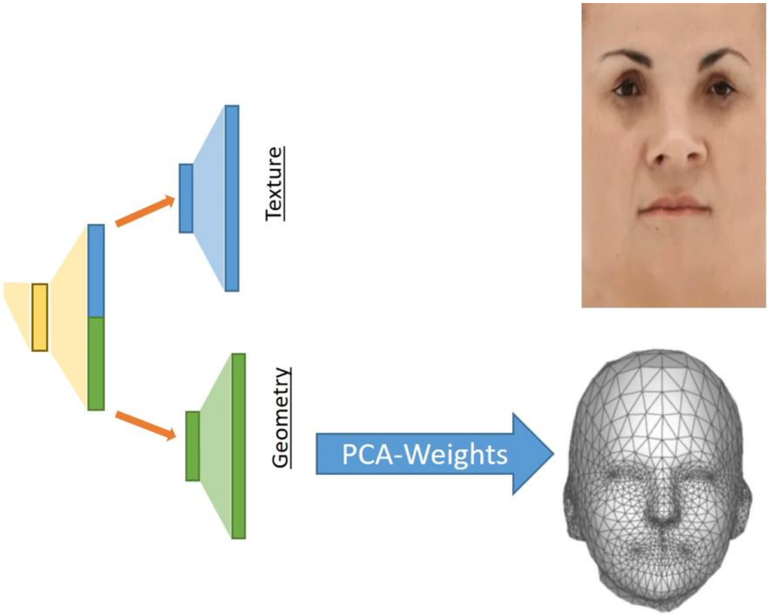
W. Paier et al., IEEE CGA 2021

Neural Face Representation

Variational Autoencoder

Latent vector

= animation parameters



W. Paier et al., IEEE CGA 2021

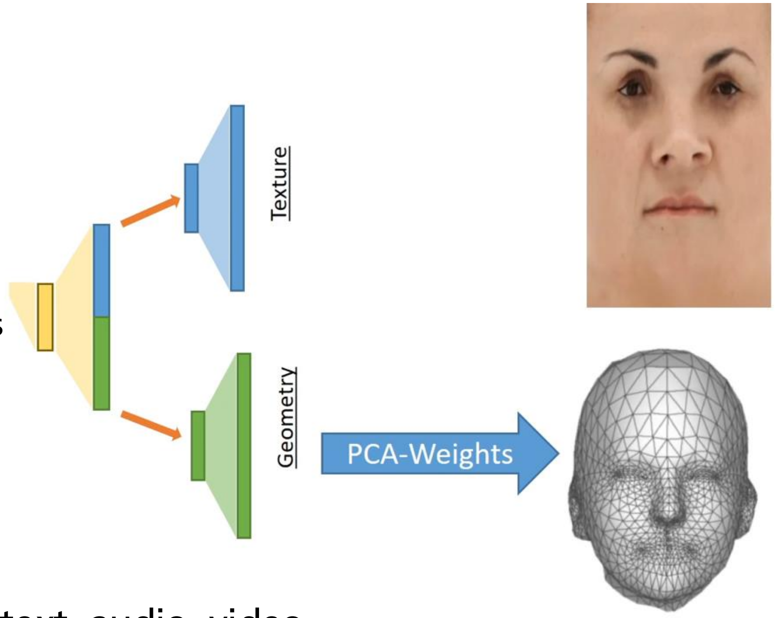
Neural Face Representation

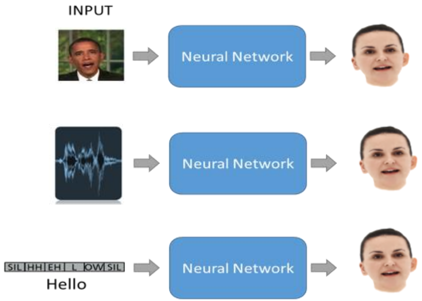
Variational Autoencoder

Latent vector

= animation parameters

Meaning ?





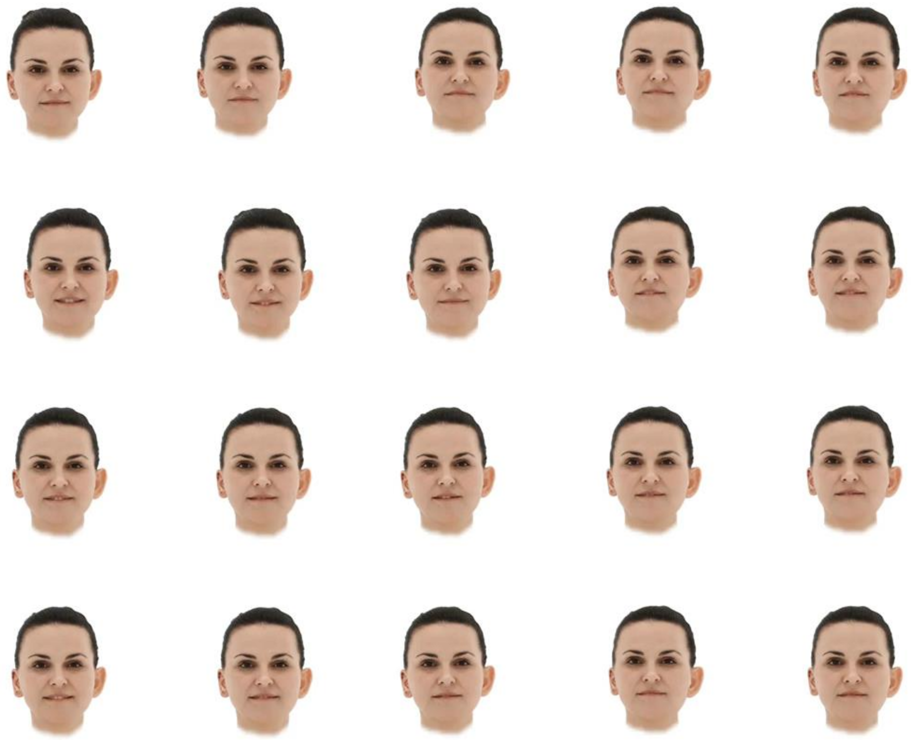
- ➢ Learn a mapping to text, audio, video

W. Paier et al., IEEE CGA 2021
W. Paier et al., Graphics & Visual Computing

Neural Face Animation
Speech synthesis results

Neural Face Animation

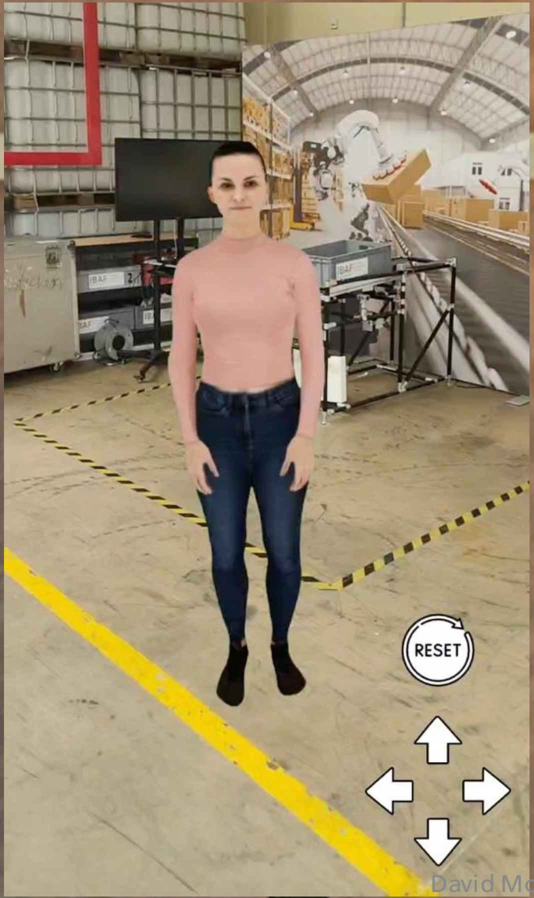
Speech synthesis results – different styles

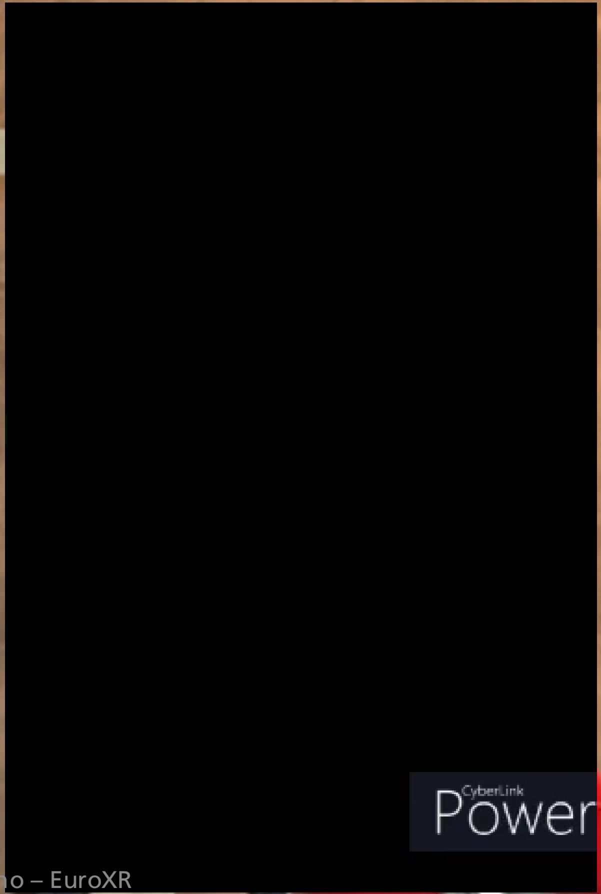


Animation Control

Driving video Style 1 Style 2 Style 3

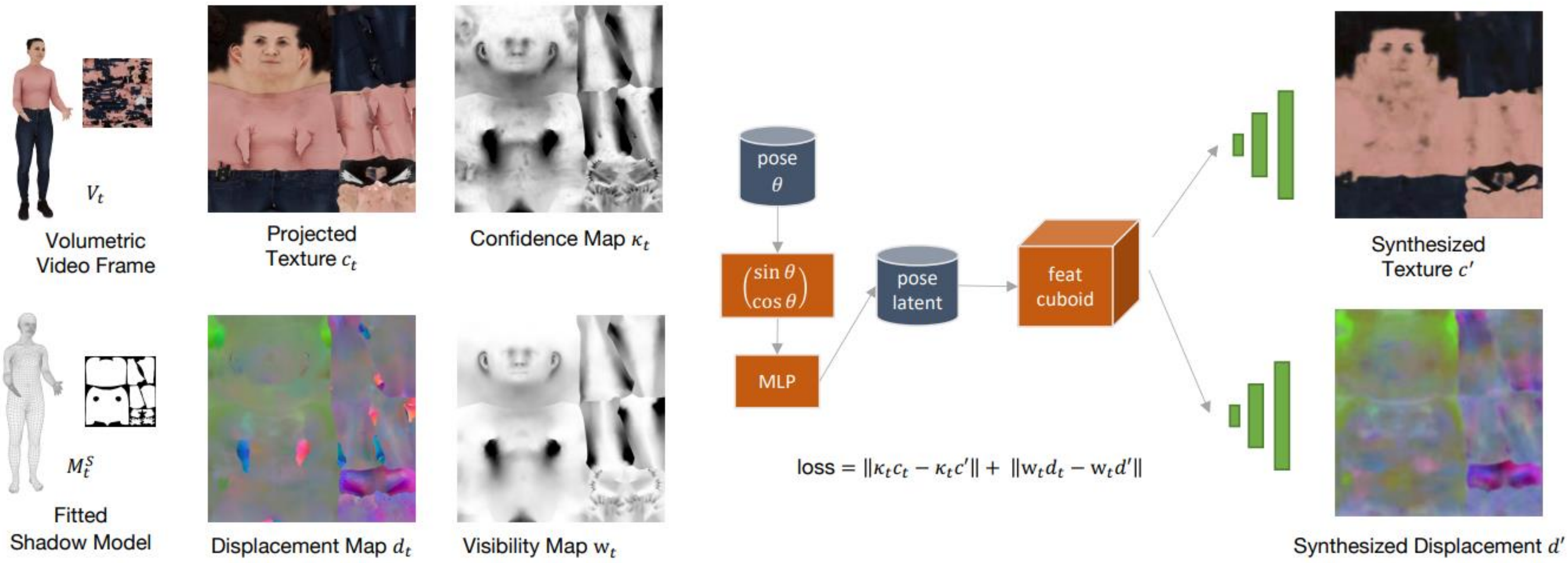
Integrated Avatars





Learning a Generative Model for Virtual Humans

Training a pose-dependent decoder



Morgenstern et al., IEEE VR 2024

Learning a Generative Model for Virtual Humans

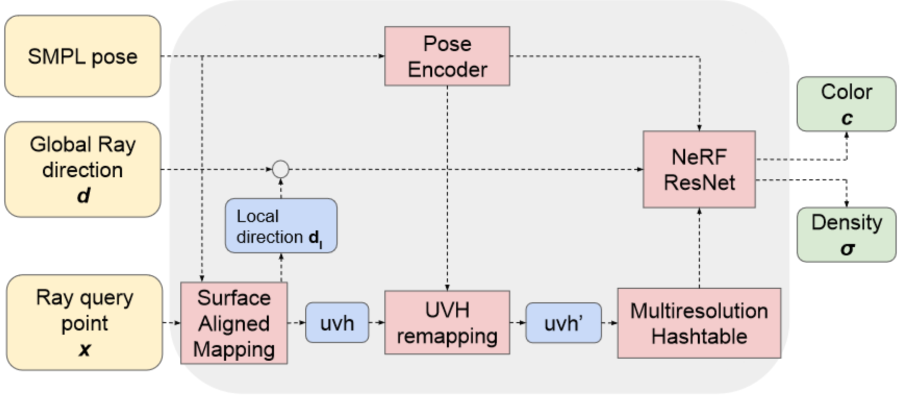
Results

synthesized
model

volumetric
video

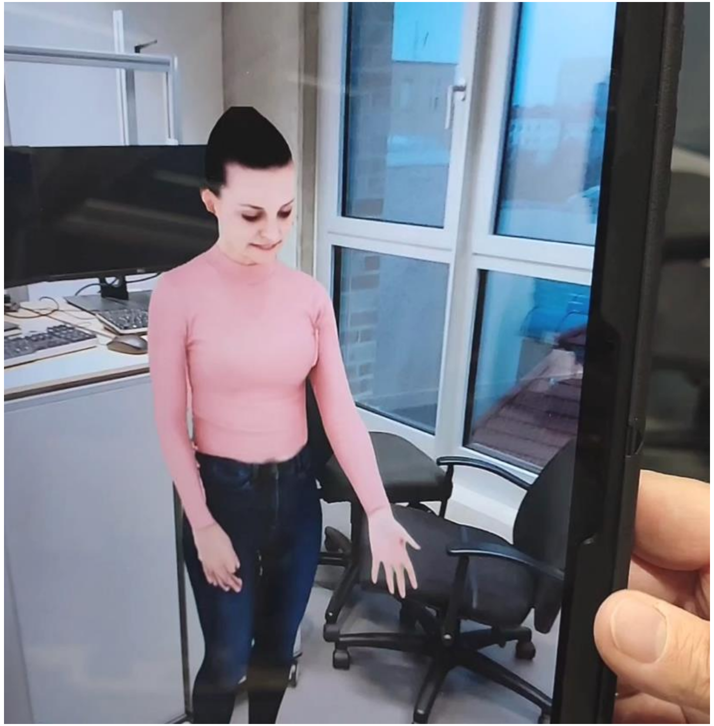
SMPL Aligned NeRF

[Knoll VISAPP 2024]



- Extended Surface aligned NeRF [XFM22]
- Mapping of canonical space to uv-height coordinates of SMPL model
- Remapping of coordinates with learned frame embeddings to account for alignment errors
- Can be animated via skeleton to unseen poses

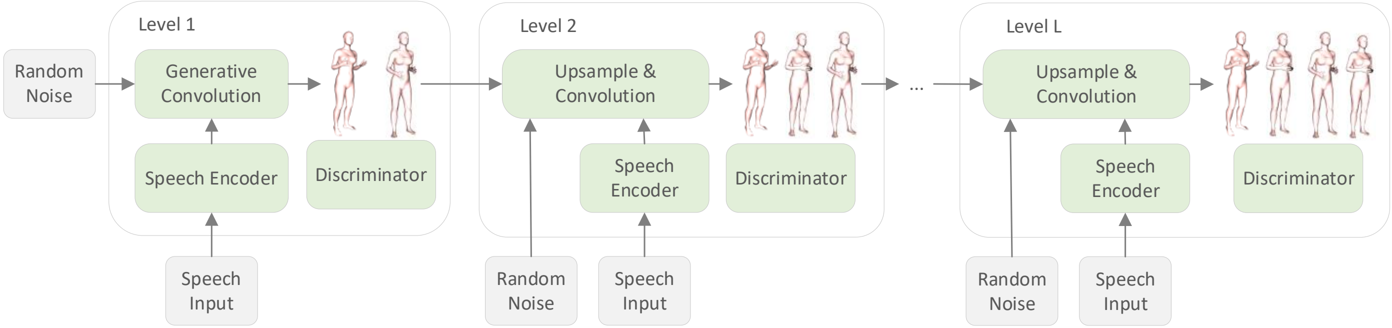
Learning a Generative Model for Virtual Humans
AR Visualization



[Morgenstern IEEE VR 2024]

Human Animation

Learning Co-Speech Gesture Generation

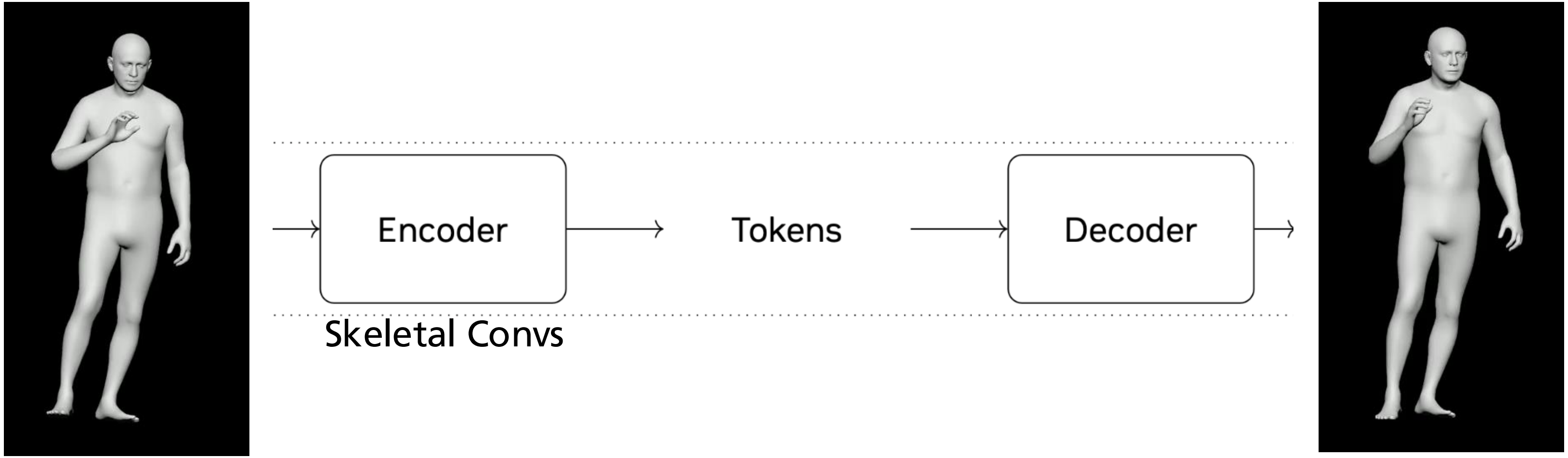


Moreno et al., CVMP 2024

Gesture Animation

New approach

- We include skeletal-aware convolution layers in different motion tokenizers
- Each motion tokenizer learns a compact motion representation of the upper body, lower body and hands



Gesture Animation

Tokenized Motion

- A transformer network predicts motion tokens based on semantic text embeddings (Llama embeddings)

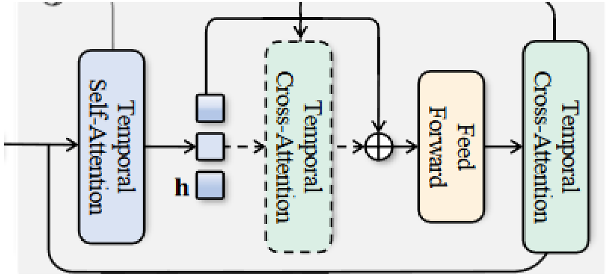
Text

(Llama embeddings)

Masked

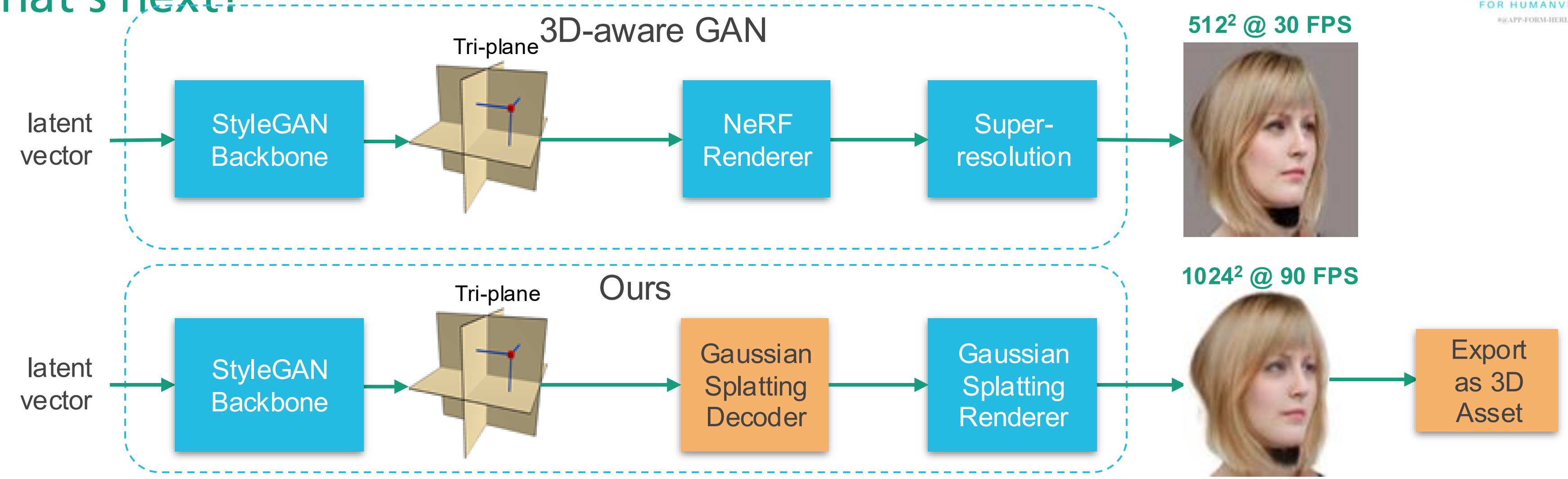
Motion Tokens

(upper/lower/hands)



From ~~Volumetric Video~~ to Realistic Avatars

What's next?



- Gaussian Splatting Renderer, standard rendering possible, e.g. Unity, 150fps
- Novel mapping network synthesizing Gaussian properties
- End to end training without superresolution network

[Barthel CVPR 2024]

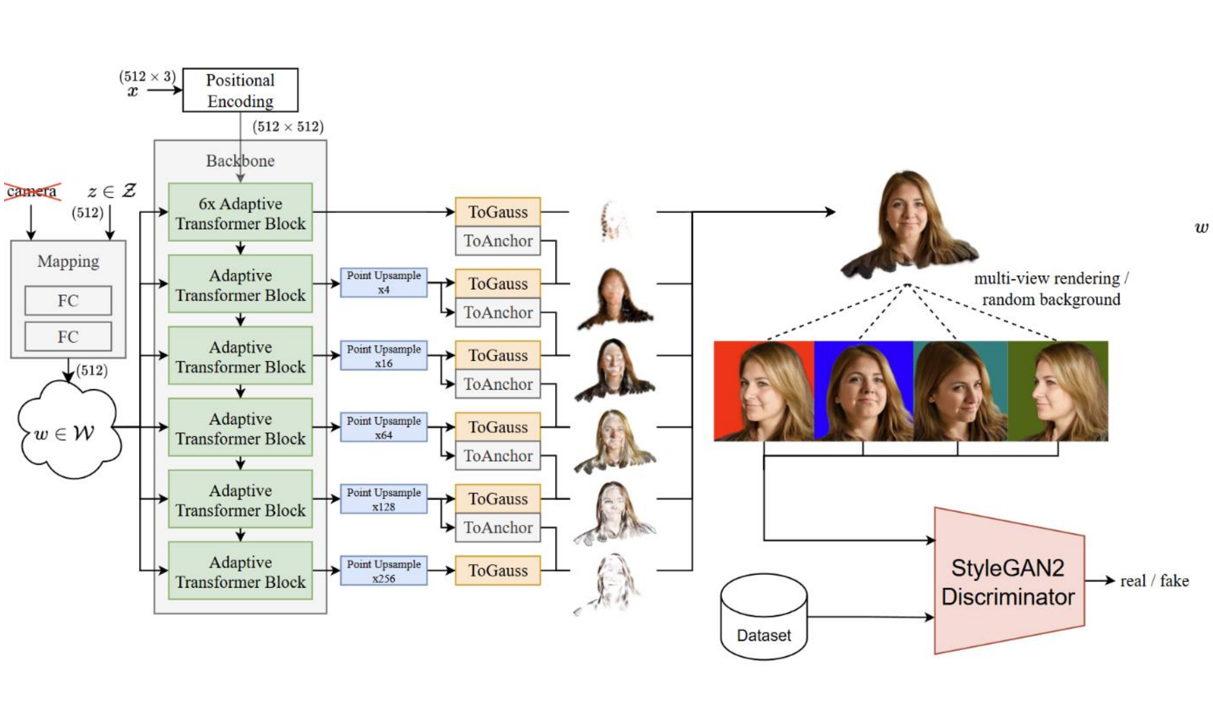
From ~~Volumetric Video~~ to Realistic Avatars

What's next?

Peter Eisert

From ~~Volumetric Video~~ to Realistic Avatars

What's next?



F. Barthel et al., Arxiv 2025

- Gaussian Splatting representation for high quality and real-time rendering
- Easy personalization by GAN structure
- synthesizes arbitrary faces
- GAN inversion adjusts model to 1 or more pictures of a person

Submitted to NeurIPS

F. Barthel et al., Arxiv 2025

Editing of Face Properties Through Latent Space Offsets

David Moreno – EuroXR

Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, HHI

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE

WE PUT SCIENCE INTO ACTION.

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

SPONSORED BY THE
Federal Ministry of Education and Research

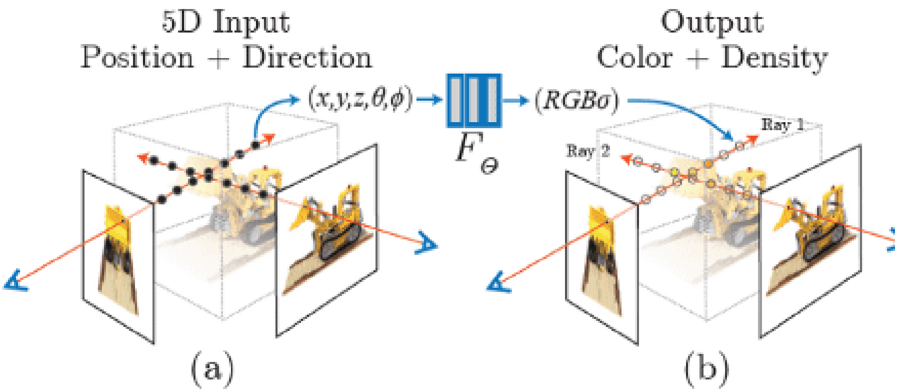
Contact:

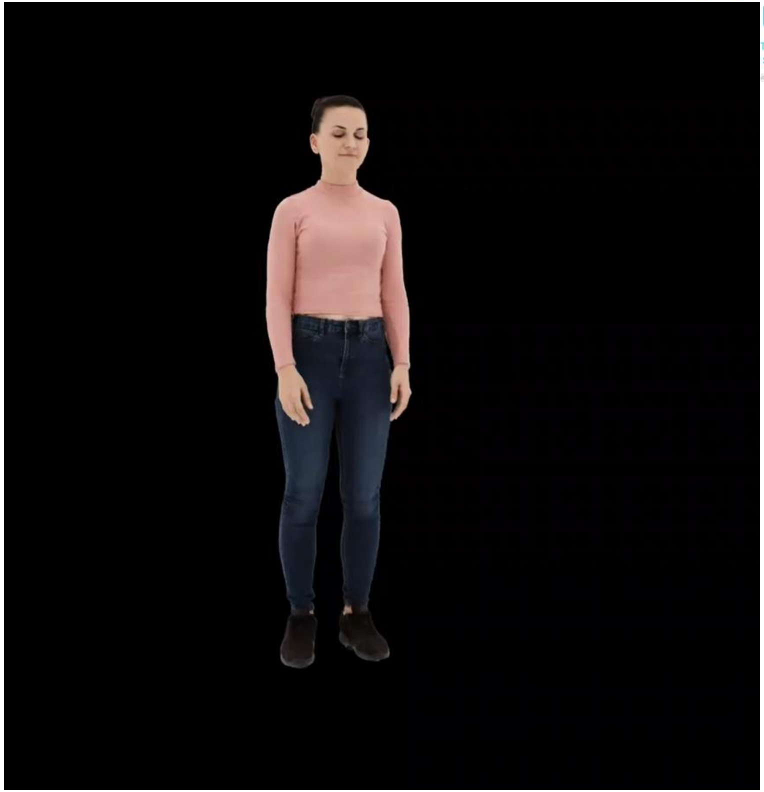
David Moreno
david.moreno@hhi.fraunhofer.de

Peter Eisert
peter.eisert@hhi.fraunhofer.de

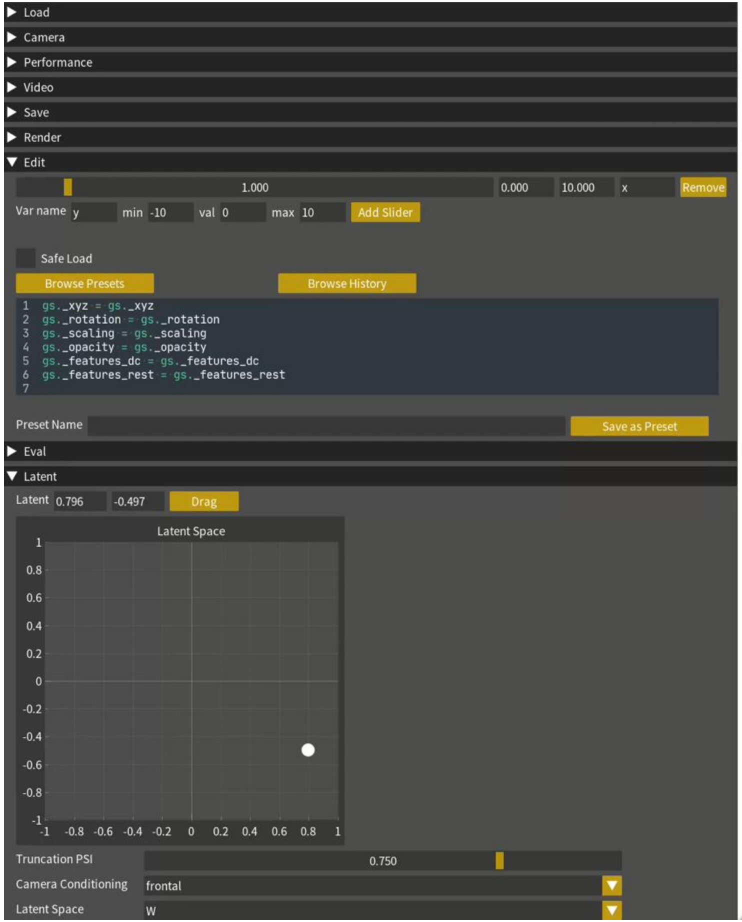
Einsteinufer 37
10587 Berlin

Neural Radiance Fields (NeRF)





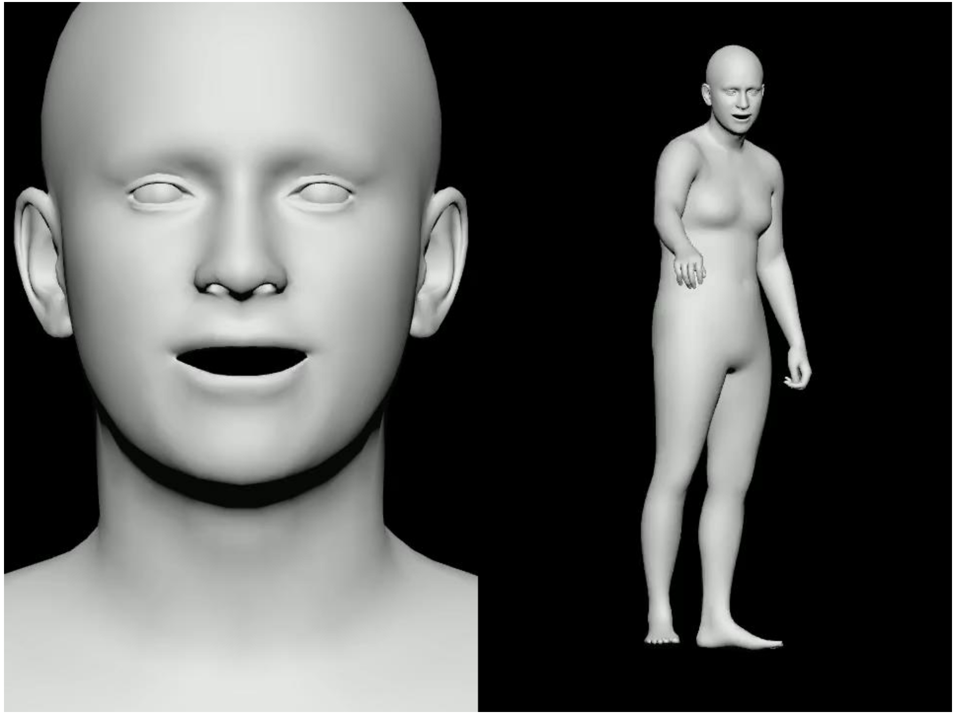
- MLP that reconstructs RGBA for a 3D point x
- Volume rendering by accumulating line of sight
- Enables static 3D object rendering
- Instant NGP for fast training



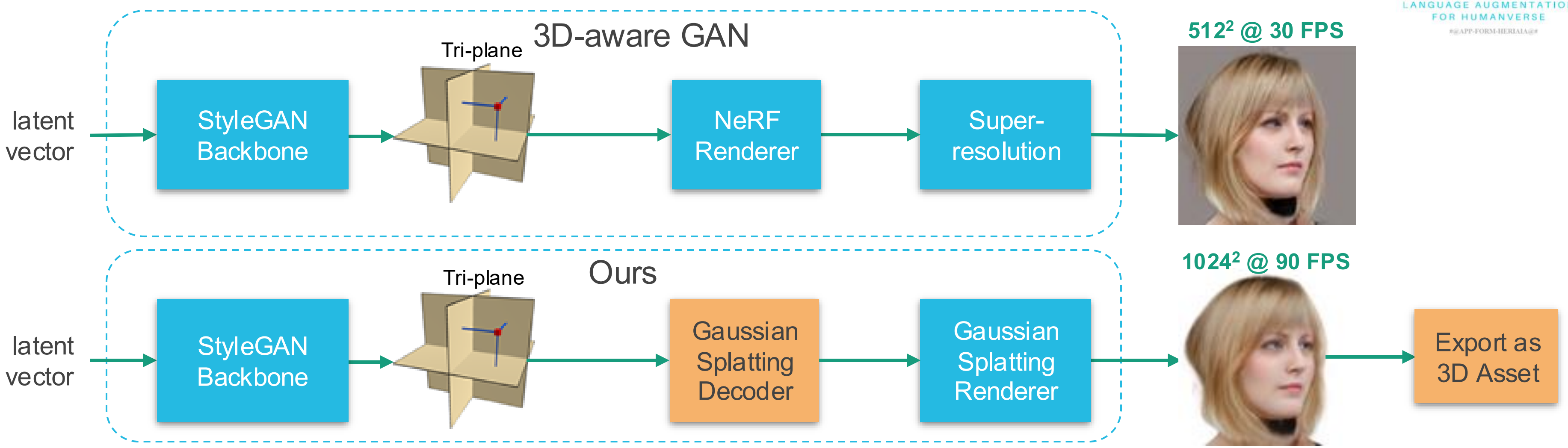
F. Barthel et al., Arxiv 2025

Fraunhofer HHI

Gesture Animation – Preliminary Results



GAN Synthesis of Gaussian Splatting Faces



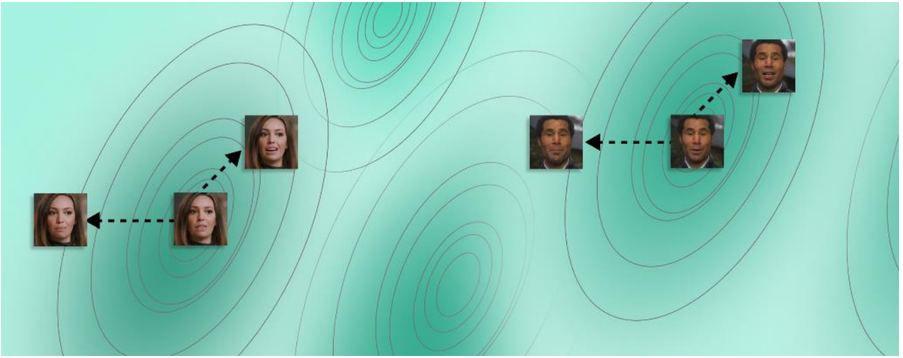
- Presented @ CVPR 2024
- Mapping network that creates properties of Gaussian Splats (size, color, opacity, orient.)
- Fast rendering with standard libraries
- No need for super-resolution step

Editable 3D Renderings

Fraunhofer HHI

Lip Animation Through GAN Inversion

- GAN inversion for frames of a video sequence
- Create latent vectors defining motion space
- For different individuals, mimic directions can differ in latent space
- Retraining of generator to reshape GAN space such that same motion of different individuals is aligned

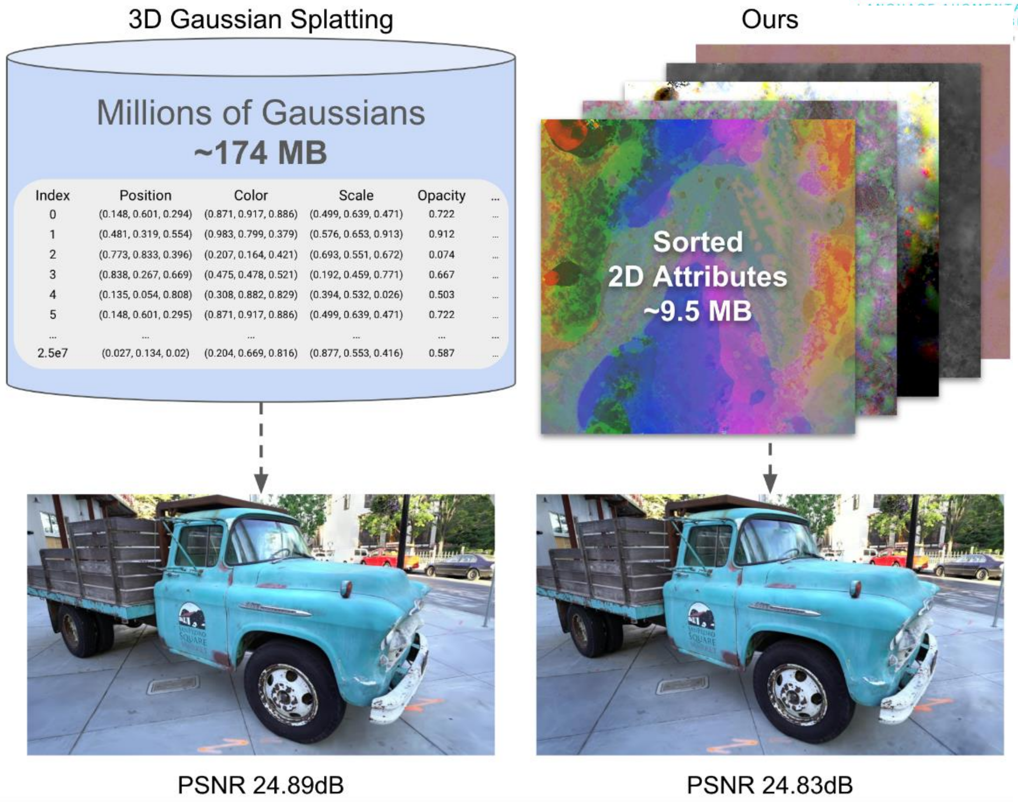


Expression Transfer

Compression of Gaussian Splatting Scenes

- Gaussian Splatting scenes show high quality results but are very large in size
- Efficient coding and storage necessary
- Our codec compresses exploiting
  - high dimensional sorting
  - scene refinement
- Size reduction of a factor 15-25 at almost same quality

[Morgenstern et al. ECCV 2024]



Index	Position	Color	Scale	Opacity	...
0	(0.148, 0.601, 0.294)	(0.871, 0.917, 0.886)	(0.499, 0.639, 0.471)	0.722	...
1	(0.481, 0.319, 0.554)	(0.983, 0.799, 0.379)	(0.576, 0.653, 0.913)	0.912	...
2	(0.773, 0.833, 0.396)	(0.207, 0.164, 0.421)	(0.693, 0.551, 0.672)	0.074	...
3	(0.838, 0.267, 0.669)	(0.475, 0.478, 0.521)	(0.192, 0.459, 0.771)	0.667	...
4	(0.135, 0.054, 0.808)	(0.308, 0.882, 0.829)	(0.394, 0.532, 0.026)	0.503	...
5	(0.148, 0.601, 0.295)	(0.871, 0.917, 0.886)	(0.499, 0.639, 0.471)	0.722	...
...	...	...	...	...	...
2.5e7	(0.027, 0.134, 0.02)	(0.204, 0.669, 0.816)	(0.877, 0.553, 0.416)	0.587	...

Fraunhofer-Institut für Nachrichtentechnik, Heinrich-Hertz-Institut, HHI

WE PUT SCIENCE INTO ACTION.

Contact:

Peter Eisert
peter.eisert@hhi.fraunhofer.de
+49 30 31002 614

Einsteinufer 37
10587 Berlin