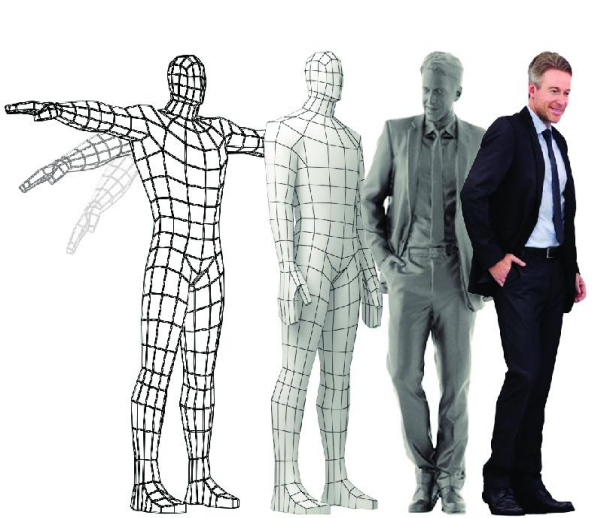
RESPONSIVE REALISTIC VIRTUAL AVATARS FOR THE METAVERSE

Peter Eisert, Fraunhofer HHI / Humboldt University Berlin



Virtual Humans as Intuitive Interfaces

Applications

Virtual teaching

VoluProf

Entertainment

INVICTUS

Immersive communication

SPIRIT

Employee- and communication training

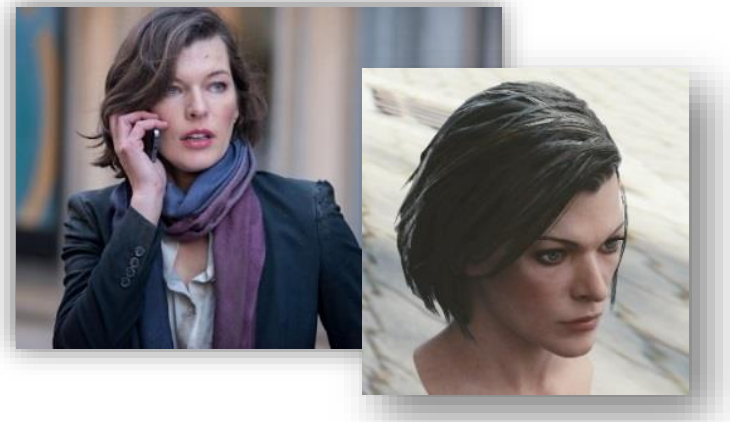
Virtual assistants

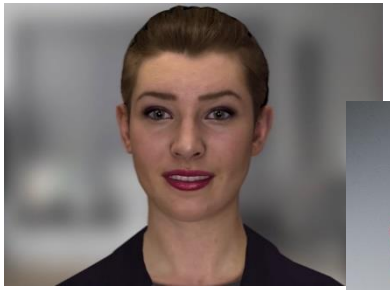
Digital sign language specialist

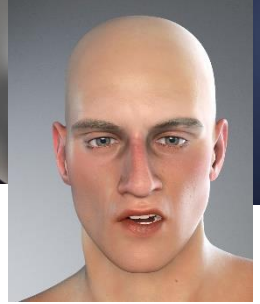
CONTENT4ALL

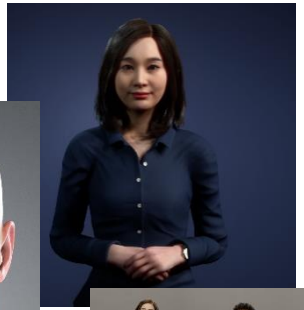
Challenges

- Realism
- Authenticity
- Feeling of virtual presence
- Feeling of connectedness



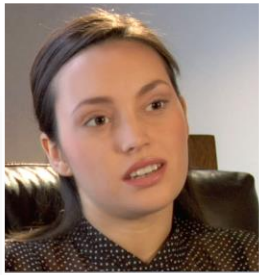






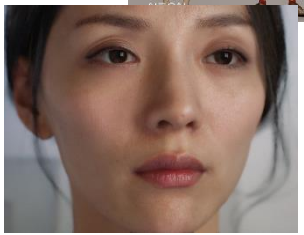
3D Scan Series

Neon



Emily, 2009

Ira, 2013



Siren, 2018

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

Volumetric Video

3D Representation with Free Viewpoint Rendering

- Typically represented by textured meshes or point clouds
- Captured by multiple cameras or depth sensors
- Free viewpoint synthesis by rendering of dynamic 3D model

Volumetric Video Studio @ Fraunhofer HHI

- 32 cameras with 20 MPixels
- 150 light panels
  - matting
  - creation of arbitrary illumination
- 2 Tb/min of raw data

Reconstruction of Volumetric Video Sequences

Integration into VR Scenes

Virtual Humans Represented as NeRFs

Realtime Rendering of Video NeRFs

static NeRF

NeRF Sequence

From Volumetric Video to Virtual Humans

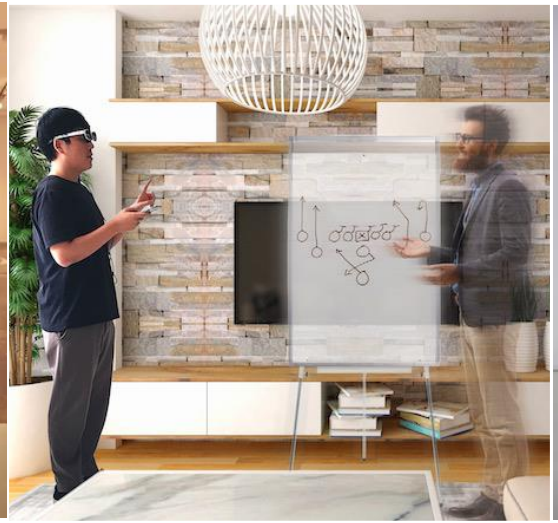
Input videos

Novel pose,
Novel viewpoint,
Novel facial expression

Magic box

Use Cases for Interactive Virtual Humans

Interactive Museum Guides



Virtual Professor/Teacher

Hybrid Events

LUMINOUS
LANGUAGE AUGMENTATION FOR HUMANVERSE
#@APP-FORM-HERIAIA@#

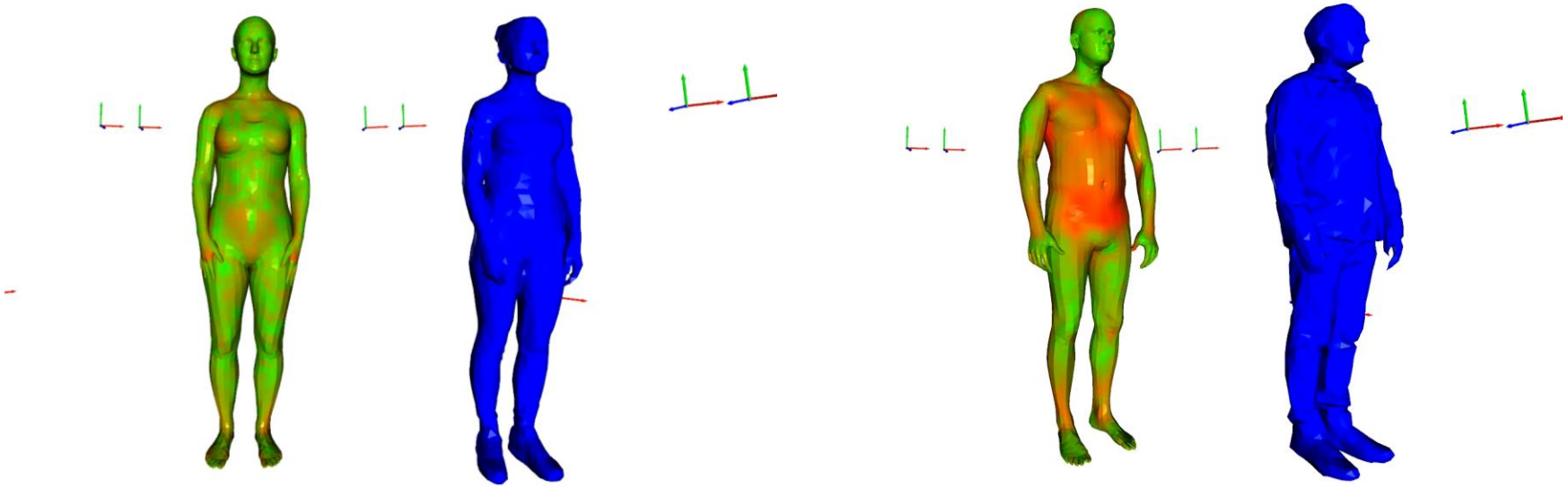
Immersive Training

INVICTUS

VoluProf

ToHyVe

Fitting SMPL for Adding Semantics



- Based on EasyMoCap (Multiview fitting)
- Extension of shape alignment of SMPL model to volumetric video

Kinematic Animation of Volumetric Video

Animating the Real World

- Kinematic animation of scan data
- Generation of new poses
- Example: interactive gaze correction
- Standardized within MPEG

original

modified head pose

A. Hilsmann et al., IET CV 2020

VR Rendering of Volumetric Video with Head Pose Correction

Holocaust Survivor Eva Umlauf

From Volumetric Video to Virtual Humans

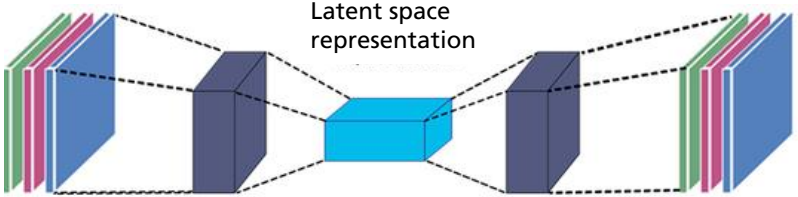
Learning Animatable Representations from Volumetric Video

Volumetric Video with fitted SMPL model

Novel pose,

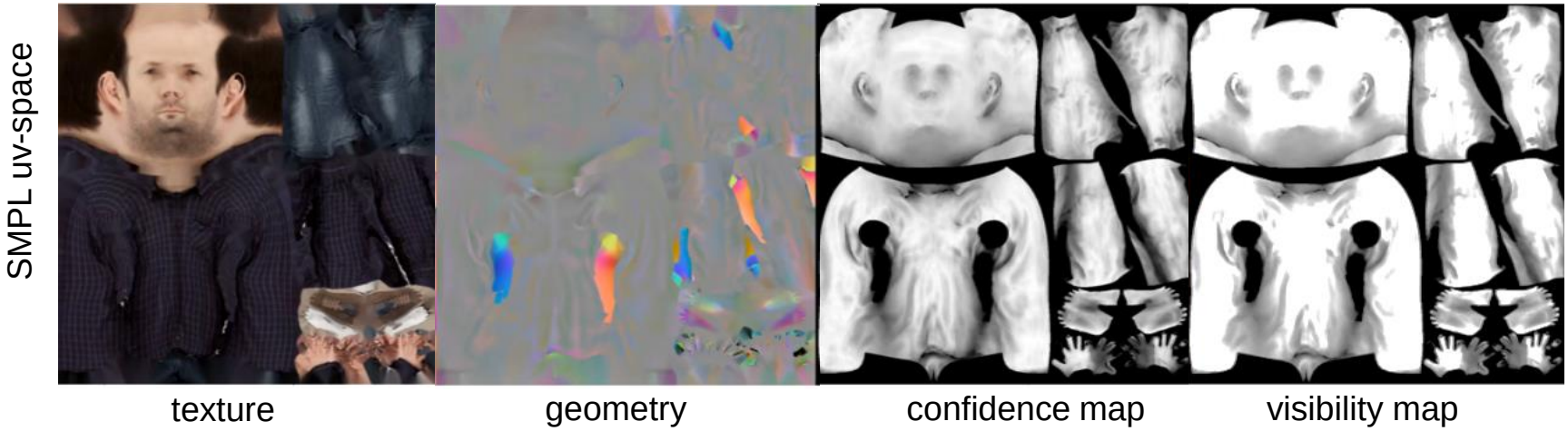
Novel viewpoint,

Novel facial expression



Learning a Generative Model for Virtual Humans

How to Represent Pose-dependent Appearance and Geometry



- Appearance : VV texture mapped to SMPL texture space
- Geometry: Displacement map (deviations between model and real geometry)
- Idea: Learn a generative model for these maps

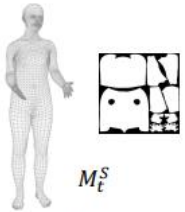
Learning a Generative Model for Virtual Humans

Training a Pose-dependent Decoder

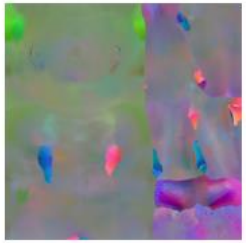
Volumetric Video Frame

Projected Texture c_t

Confidence Map κ_t

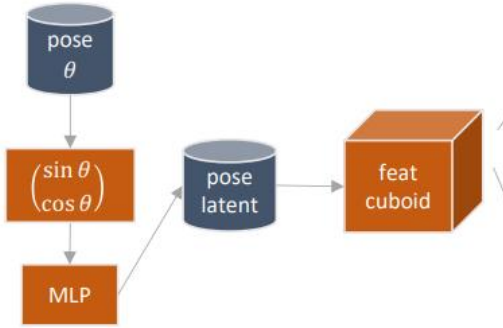


Fitted Shadow Model



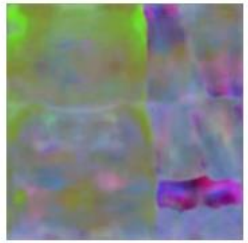
Displacement Map d_t

Visibility Map w_t



$$\text{loss} = \|\kappa_t c_t - \kappa_t c'\| + \|w_t d_t - w_t d'\|$$

Synthesized Texture c'



Synthesized Displacement d'

Learning a Generative Model for Virtual Humans

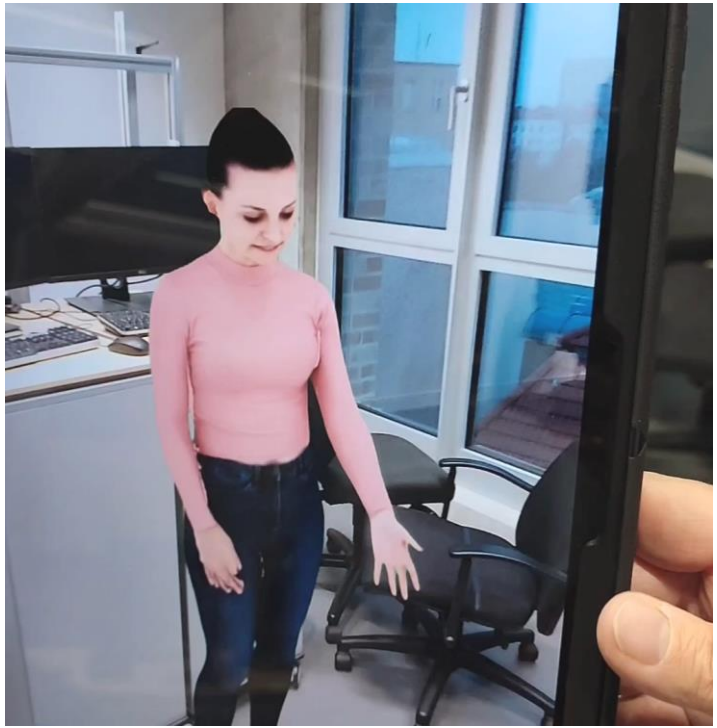
Results

synthesized model

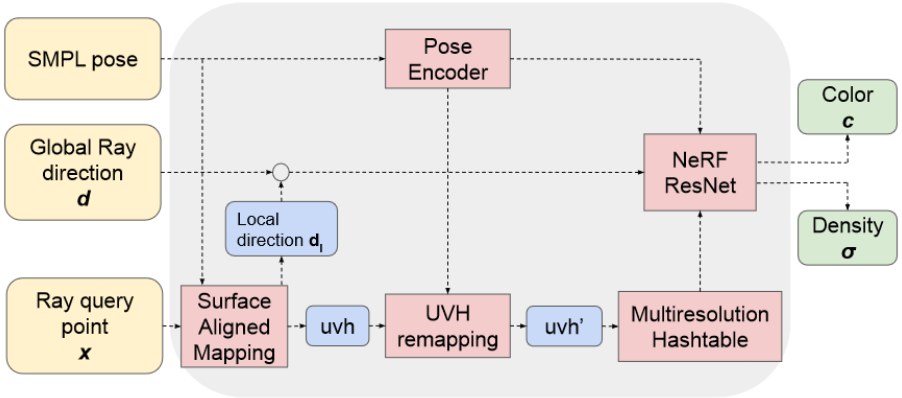
volumetric video

Learning a Generative Model for Virtual Humans

AR Visualization

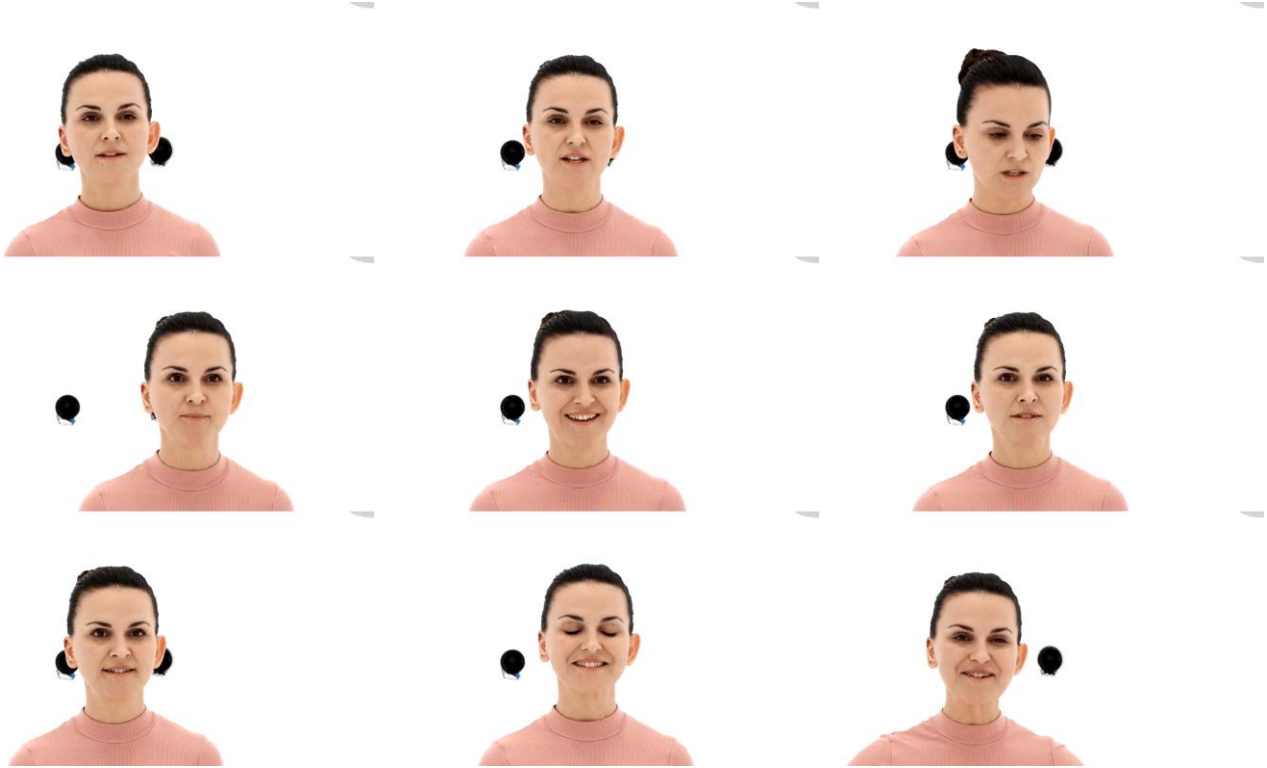


SMPL Aligned NeRF [KMH23]



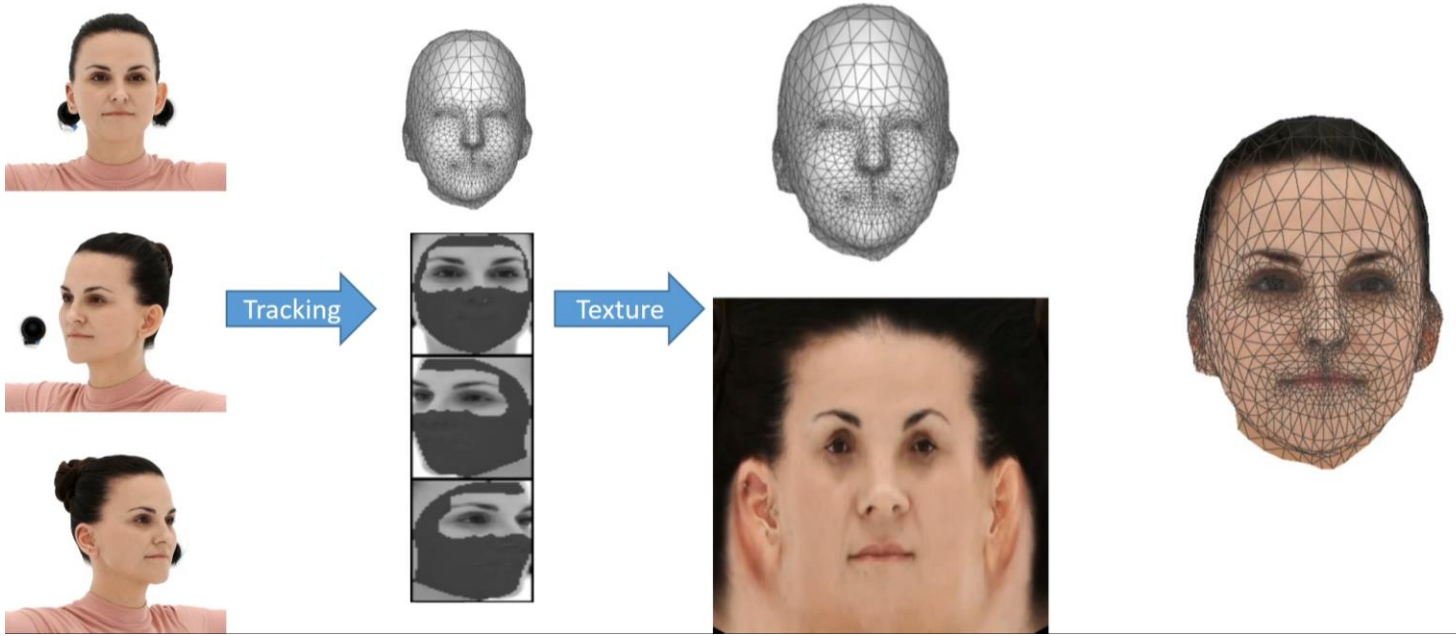
- Extended Surface aligned NeRF [XFM22]
- Mapping of canonical space to uv-height coordinates of SMPL model
- Remapping of coordinates with learned frame embeddings to account for alignment errors
- Can be animated via skeleton to unseen poses

Face Animation



Neural Face Representation

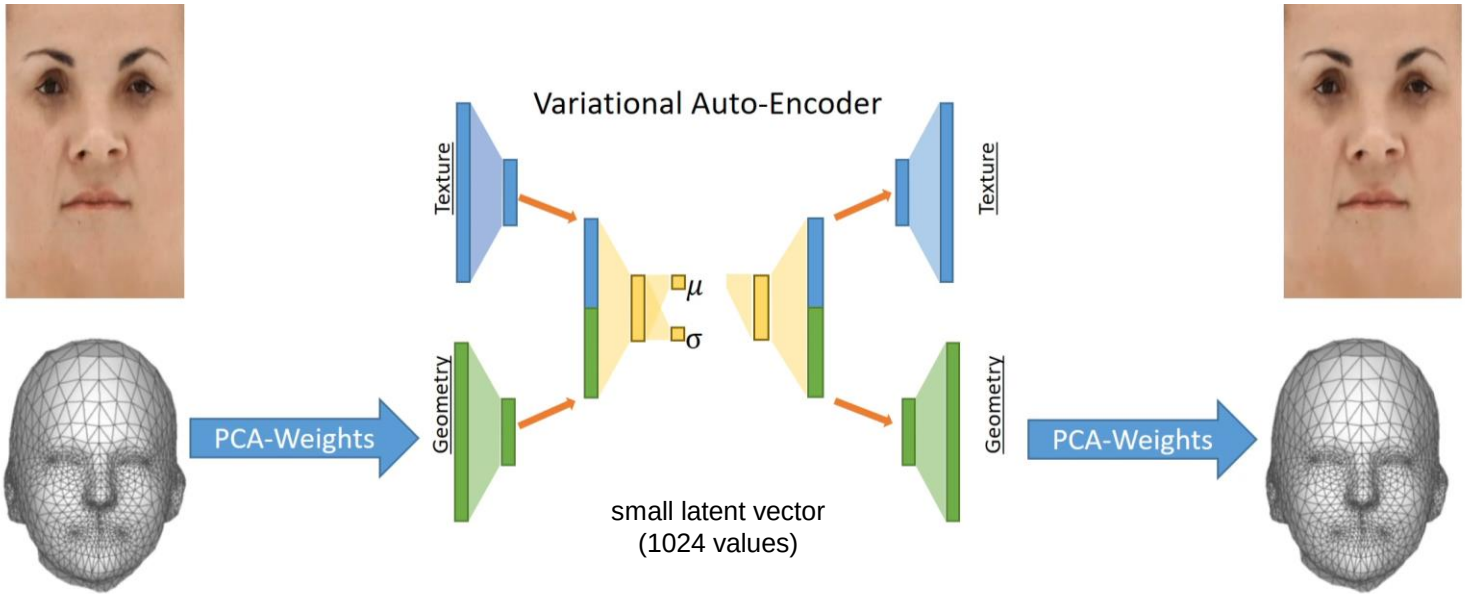
Hybrid Face Capture & Representation Learning



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

Neural Face Representation

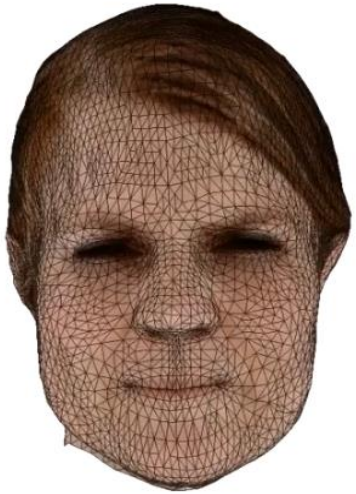
Variational Autoencoder



[W. Paier et al., IEEE CGA2021]

Neural Face Representation

Variational Autoencoder



wire-frame

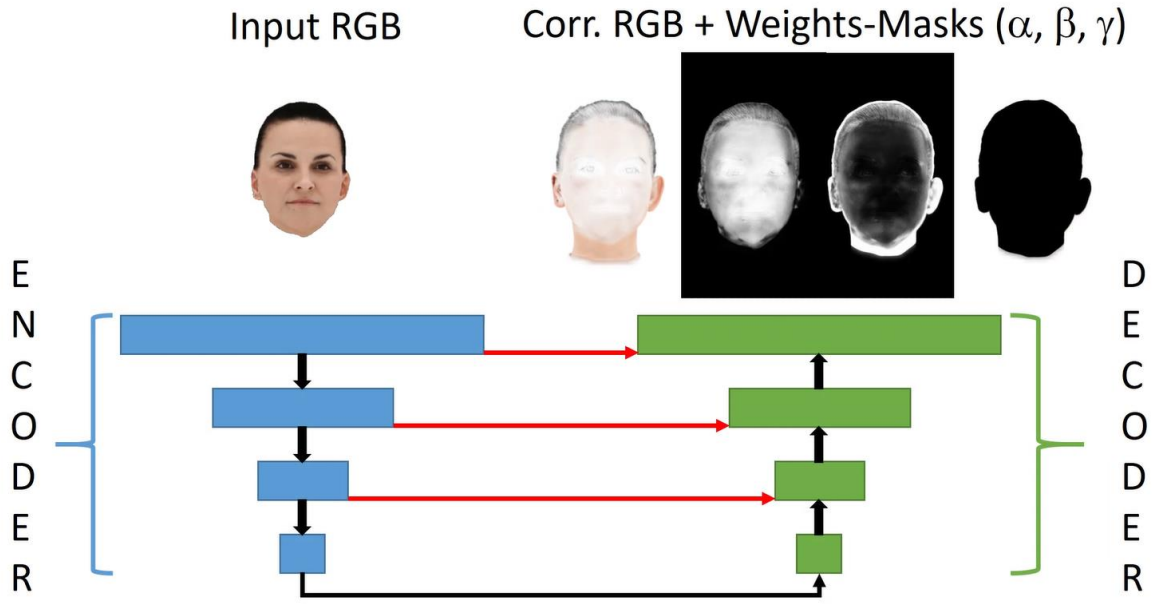
with directional light

original video

[W. Paier et al., IEEE CGA2021]

Neural Rendering

Optimizing for Photorealism



[W. Paier et al., Graphical Models]

Neural Rendering

Optimizing for Photorealism

Background

Mesh-Based Rendering

Corr. Image

Refined Rendering

[W. Paier et al., Graphical Models]

Neural Rendering

Comparison with SOTA Methods

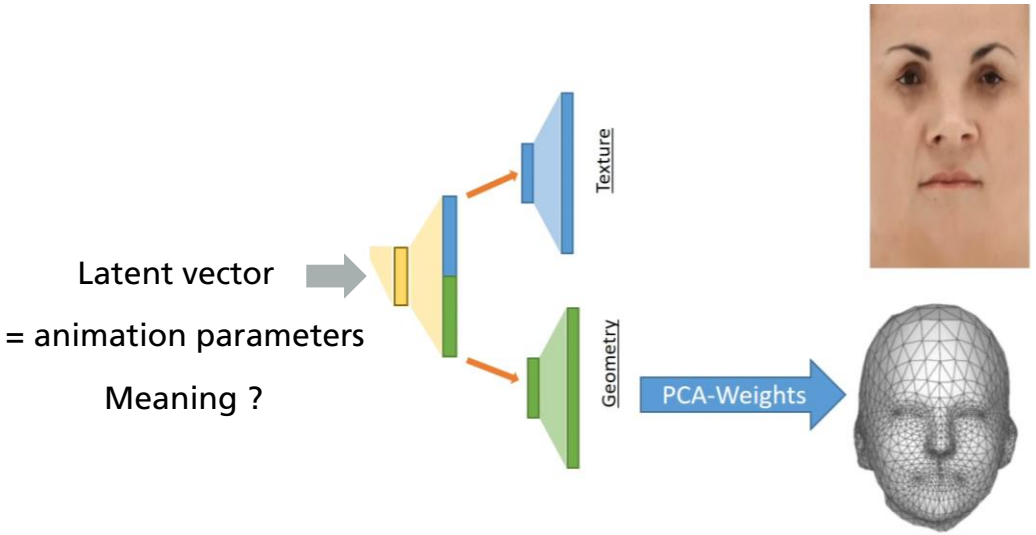
FACIAL [1] NHA [2] 4DFA [3] **OURS** **ORIGINAL**

[1] Zhang et al., Synthesizing Dynamic Talking Face with Implicit Attribute Learning, ICCV 2021
[2] Grassal et al., Neural head avatars from monocular RGB videos, CVPR 2022
[3] Gafni et al., Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction, CVPR 2021

[W. Paier et al., Graphical Models]

Neural Face Representation

Animation

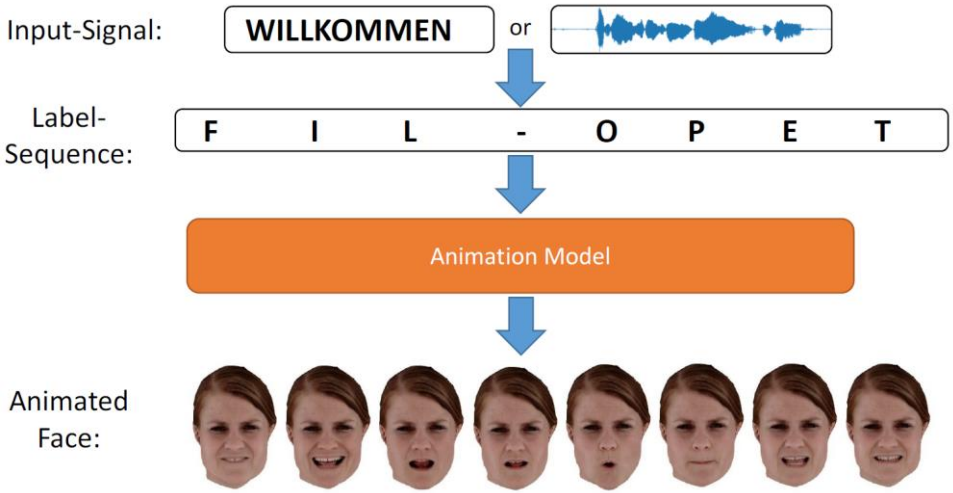


- Learn a mapping to text, audio, video

Neural Face Animation

Speech-/text-driven Facial Animation

- Animation from input text
  - text converted to viseme sequence
  - output: animation parameter sequence (frame based)
  - ambiguous mapping of speaking style and emotions
- Animation from speech
  - speech recognition (and speech alignment)
  - animation from viseme sequence
- Lip, eye and global head motion estimated from input

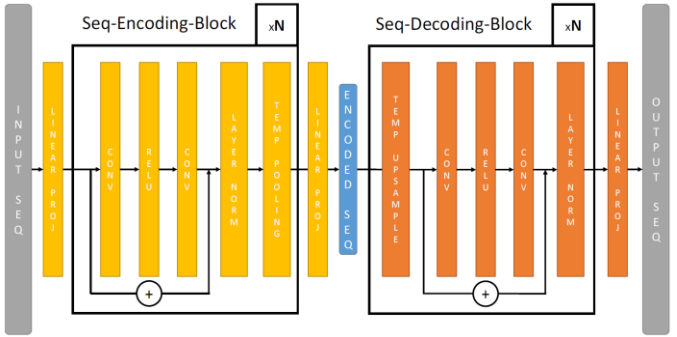


Neural Face Animation

Learning Dynamic Motion Space

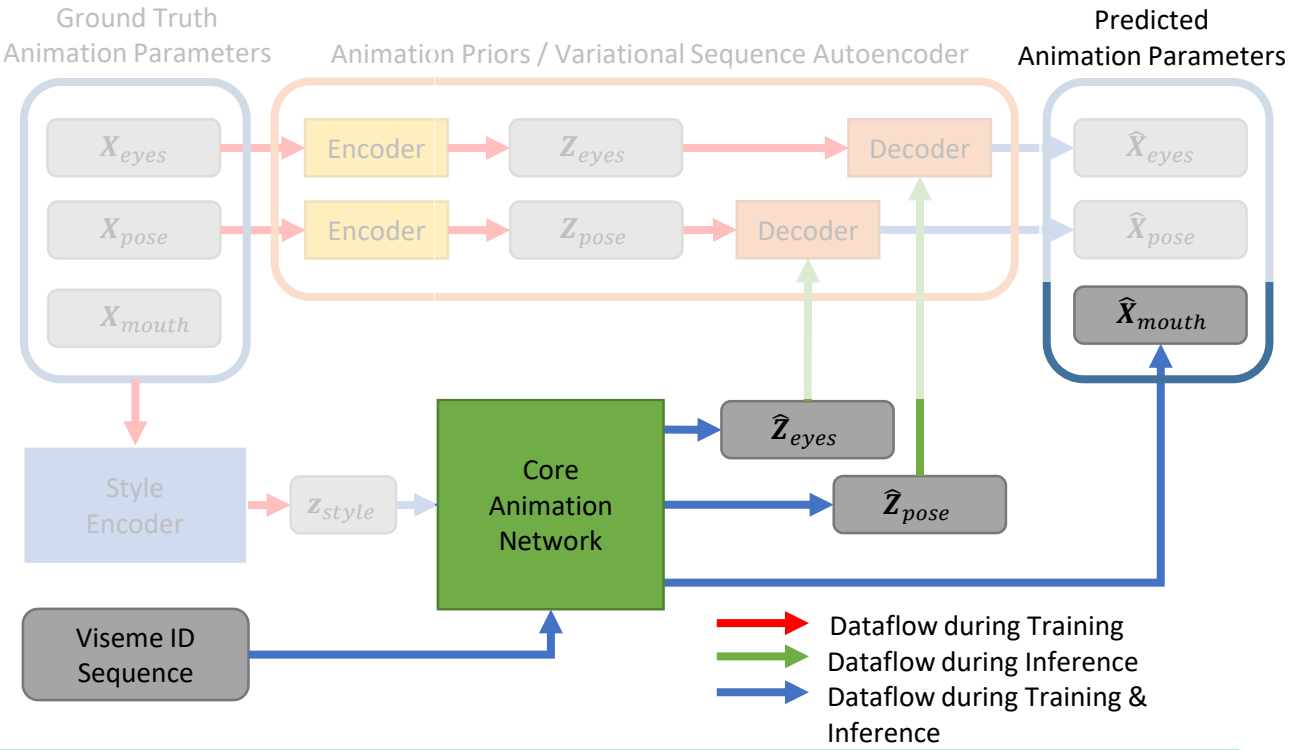
no prior | weak prior | strong prior

- Sequence to sequence learning with a variational autoencoder
- Learns plausible parameter sequences and dynamics



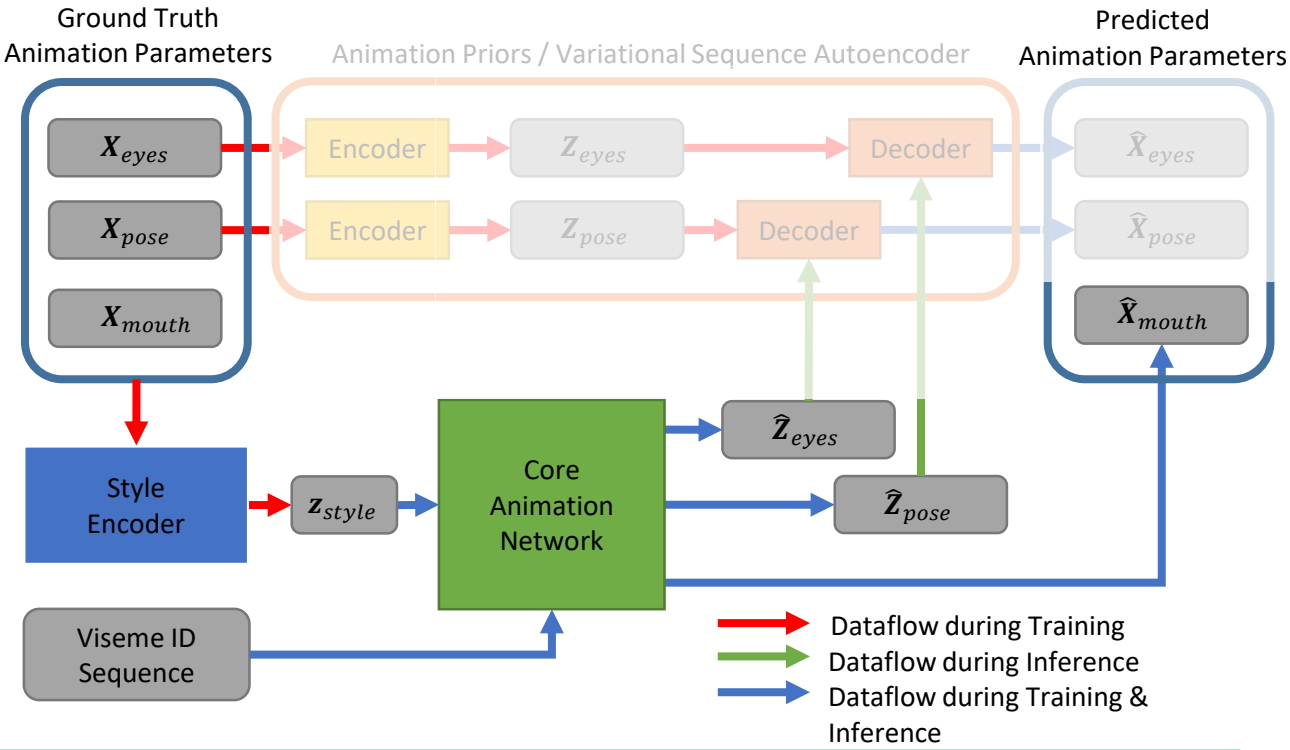
Neural Animation

- Core Animation Network
  - Viseme-IDs to animation parameters
- Style Encoder
  - Animation style space
  - Conditions core animation network
- Animation Prior
  - Regularizes eyes, head-pose



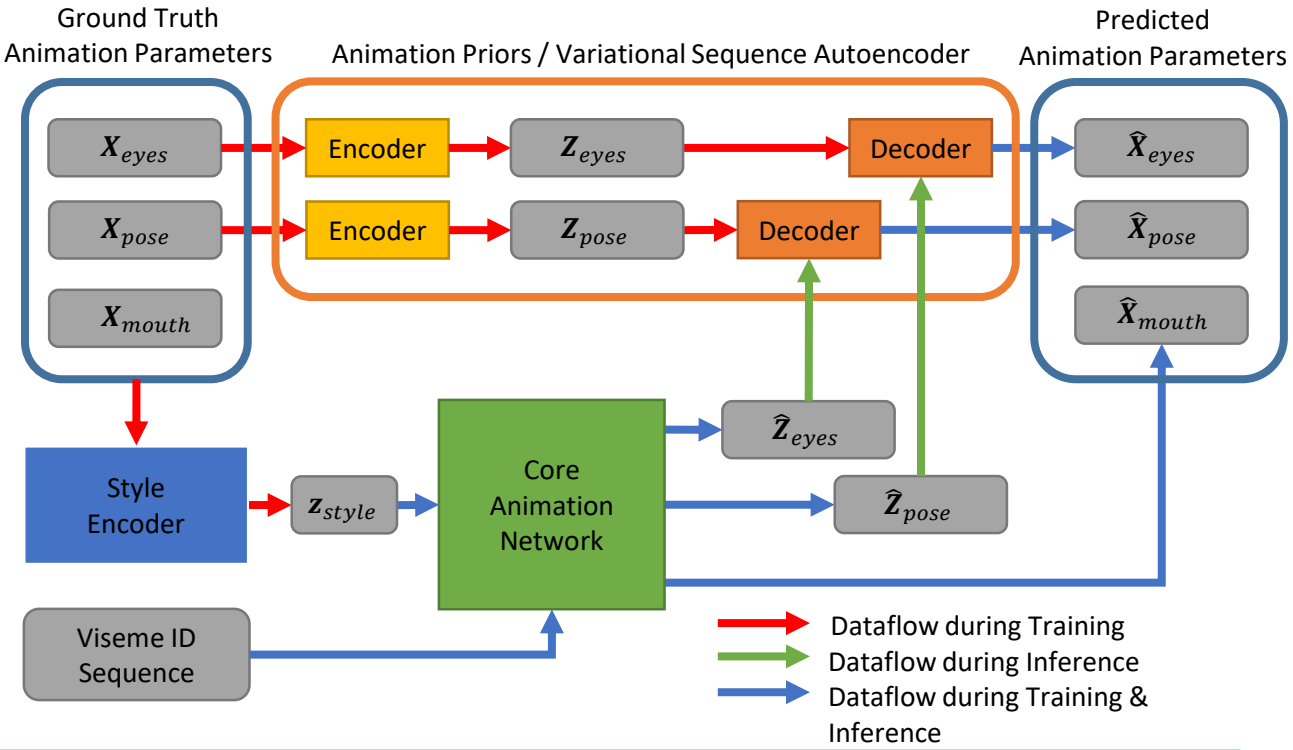
Neural Animation

- Core Animation Network
  - Viseme-IDs to animation parameters
- Style Encoder
  - Animation style space
  - Conditions core animation network
- Animation Prior
  - Regularizes eyes, head-pose

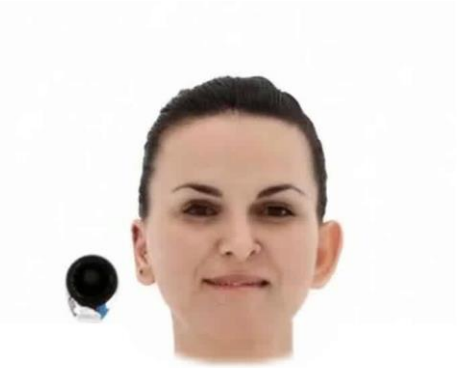


Neural Animation

- Core Animation Network
  - Viseme-IDs to animation parameters
- Style Encoder
  - Animation style space
  - Conditions core animation network
- Animation Prior
  - Regularizes eyes, head-pose

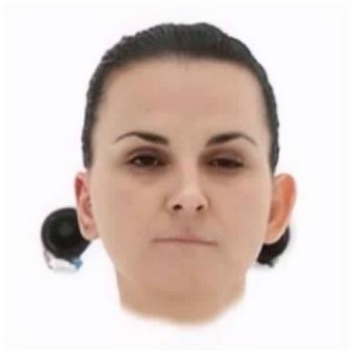


Comparison of Approaches



Facial [ZZH21]

AD-NeRF [GCL21]

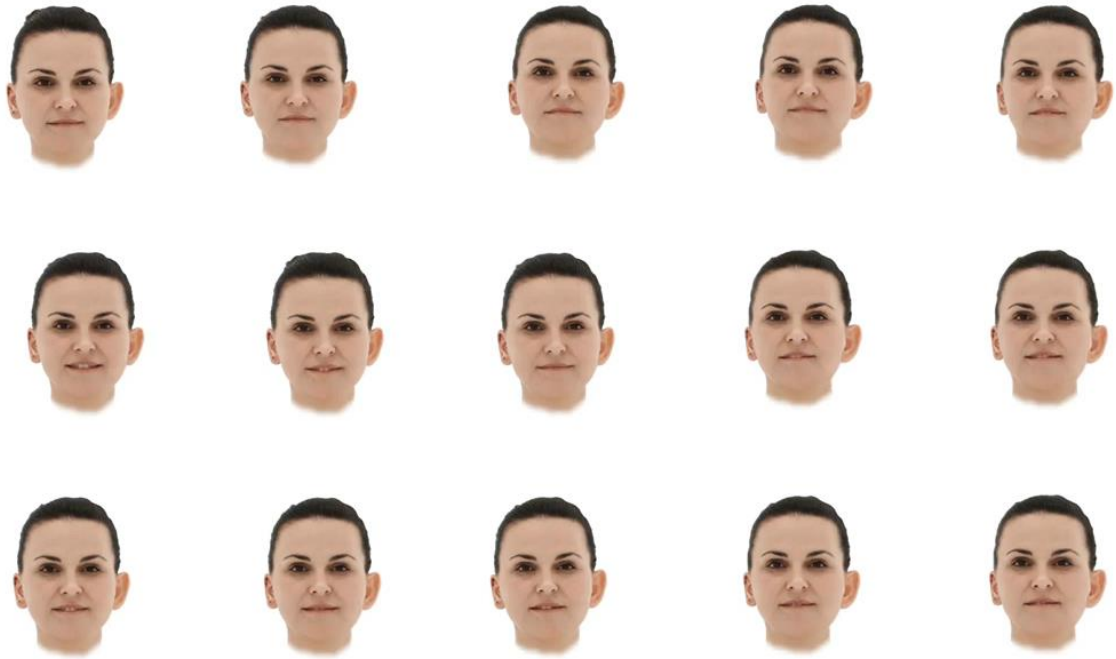


Make it Talk [ZHS20]

ours [PHE23]

Neural Face Animation

Speech Synthesis Results – Different Styles



Conclusions

- Virtual humans have gained a lot from deep learning and neural rendering
- Novel techniques like deep fakes, NeRFs, GANs, diffusion models,….
- Hybrid 3D approach
  - template model with skeleton and facial blendshapes for semantics and animation
  - neural synthesis and correction for realistic appearance
- Modelling of appearance and dynamics

- Novel applications for interactive films, games, training,….

Thanks to all the Collaborators

- Decai Chen
- Philipp Fechteler
- Ingo Feldmann
- Anna Hilsmann
- Wieland Morgenstern
- Wolfgang Paier
- Oliver Schreer
- Clemens Seibold
- Markus Worchel

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

THE FRAMEWORK PROGRAMME FOR RESEARCH AND INNOVATION
HORIZON 2020

SPONSORED BY THE
Federal Ministry of Education and Research