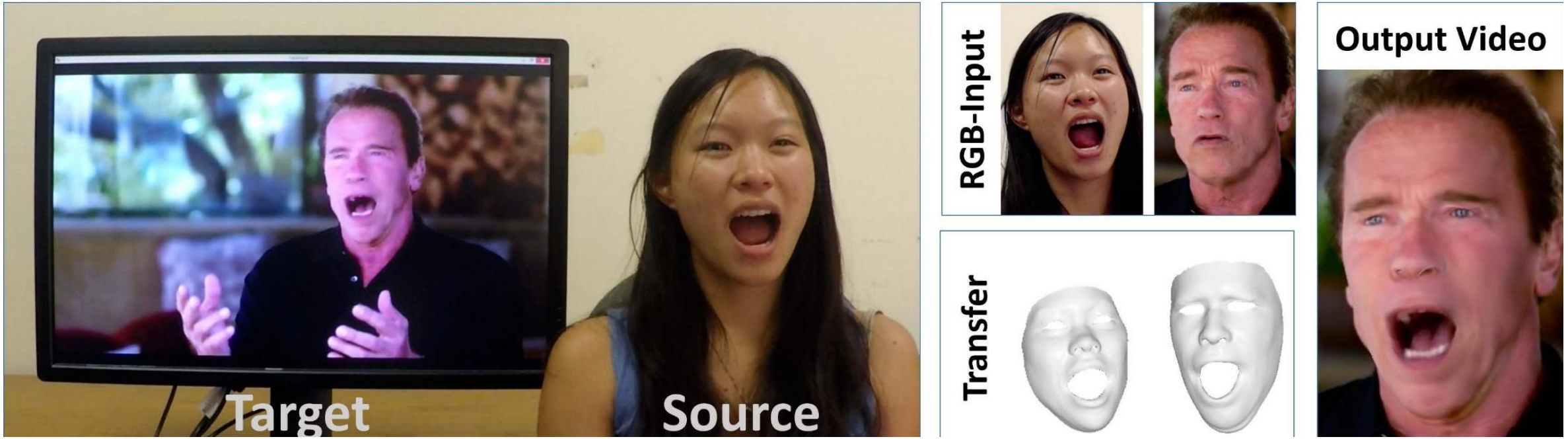
AI-Driven Video Synthesis and its Implications



Visual Computing Group
Prof. Matthias Nießner

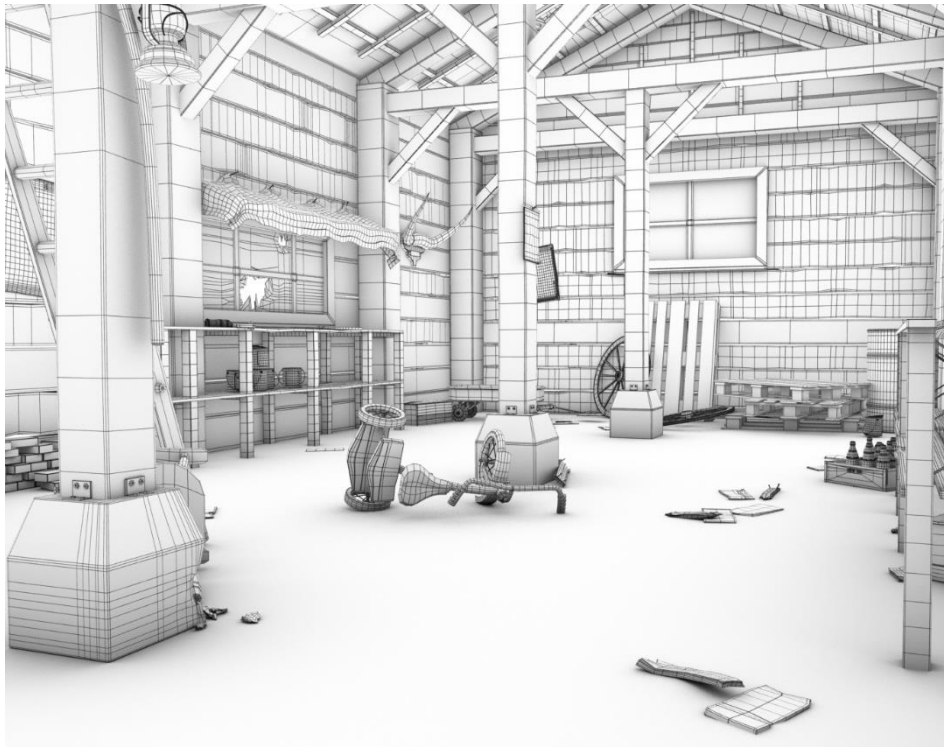
Visual Computing Group @ TUM

Photo-realistic Image Synthesis

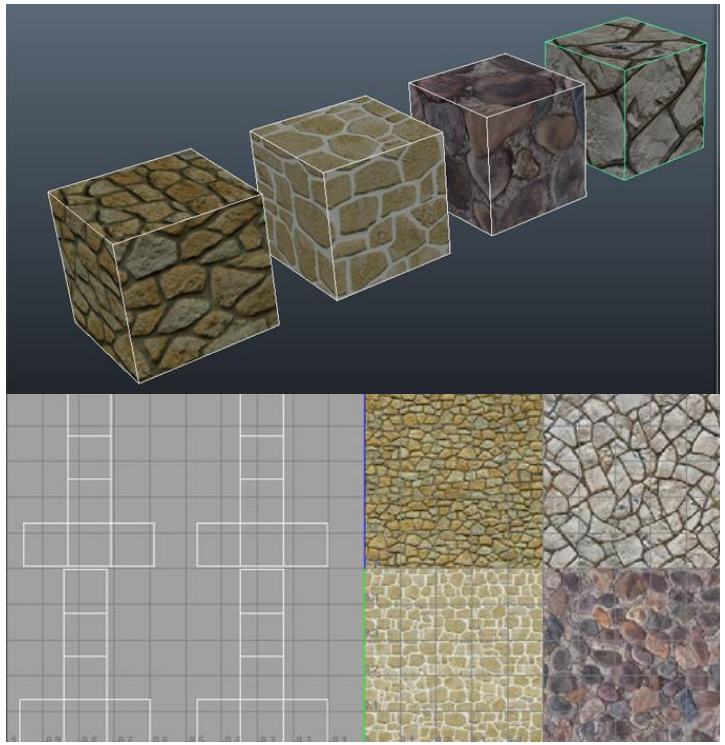
The Rendering Equation [Kajiya 86]

$$L_{\mathrm{o}}(\mathbf{x},\, \omega_{\mathrm{o}},\, \lambda,\, t) \;=\; L_e(\mathbf{x},\, \omega_{\mathrm{o}},\, \lambda,\, t) \;+\; \int_{\Omega} f_r(\mathbf{x},\, \omega_{\mathrm{i}},\, \omega_{\mathrm{o}},\, \lambda,\, t)\, L_{\mathrm{i}}(\mathbf{x},\, \omega_{\mathrm{i}},\, \lambda,\, t)\, (\omega_{\mathrm{i}} \cdot \mathbf{n})\, \mathrm{d}\,\omega_{\mathrm{i}}$$

Need 3D Content for Rendering



Geometry



Textures

Material & Lighting

Computer Vision as Inverse Graphics

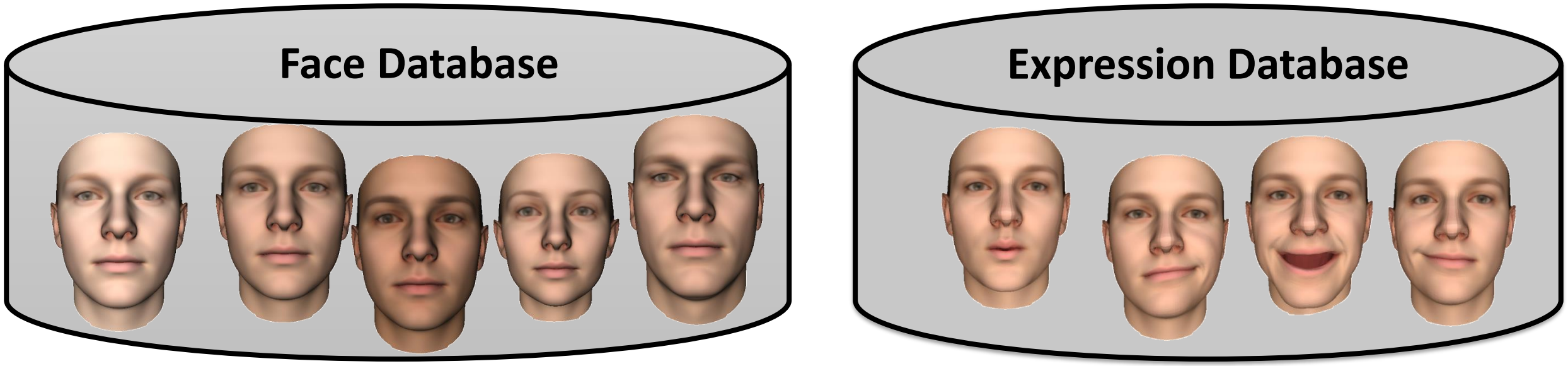
$$L_o(\mathbf{x}, \omega_o, \lambda, t)^{-1}$$

Can we invert the Rendering Equation?

Computer Vision as Inverse Graphics

$$E(P) =$$

Priors: Parametric Face Model



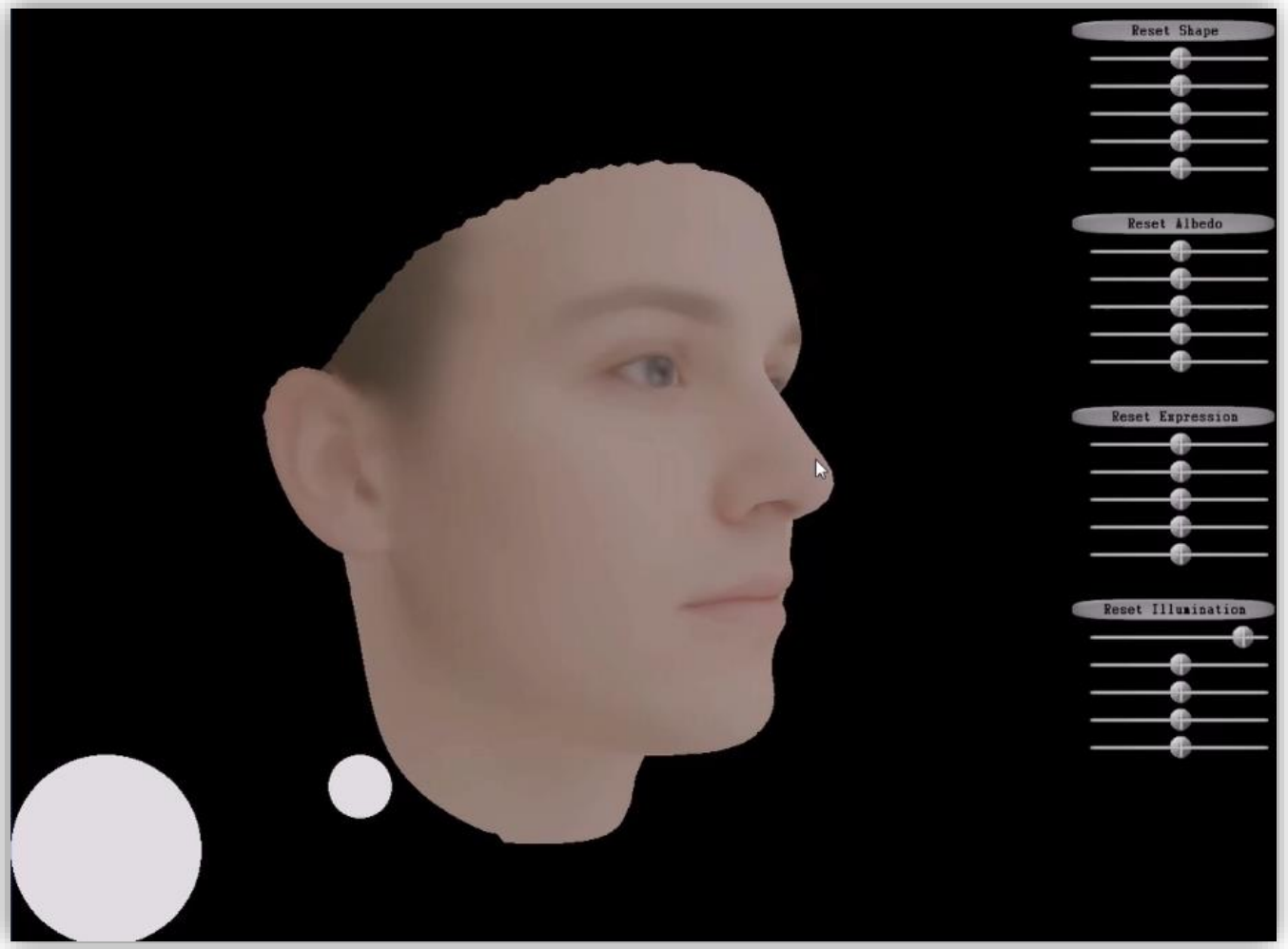
[Blanz and Vetter 99] BlendShapes
[Alexander et al. 09/10] Digital Emily
[Chen et al. 14] FaceWarehouse
…

Parametric Face Model

Rigid Pose

$$\boldsymbol{P} = \begin{pmatrix} \Phi \\ \\ \\ \end{pmatrix}$$

$|\boldsymbol{P}| = 6$



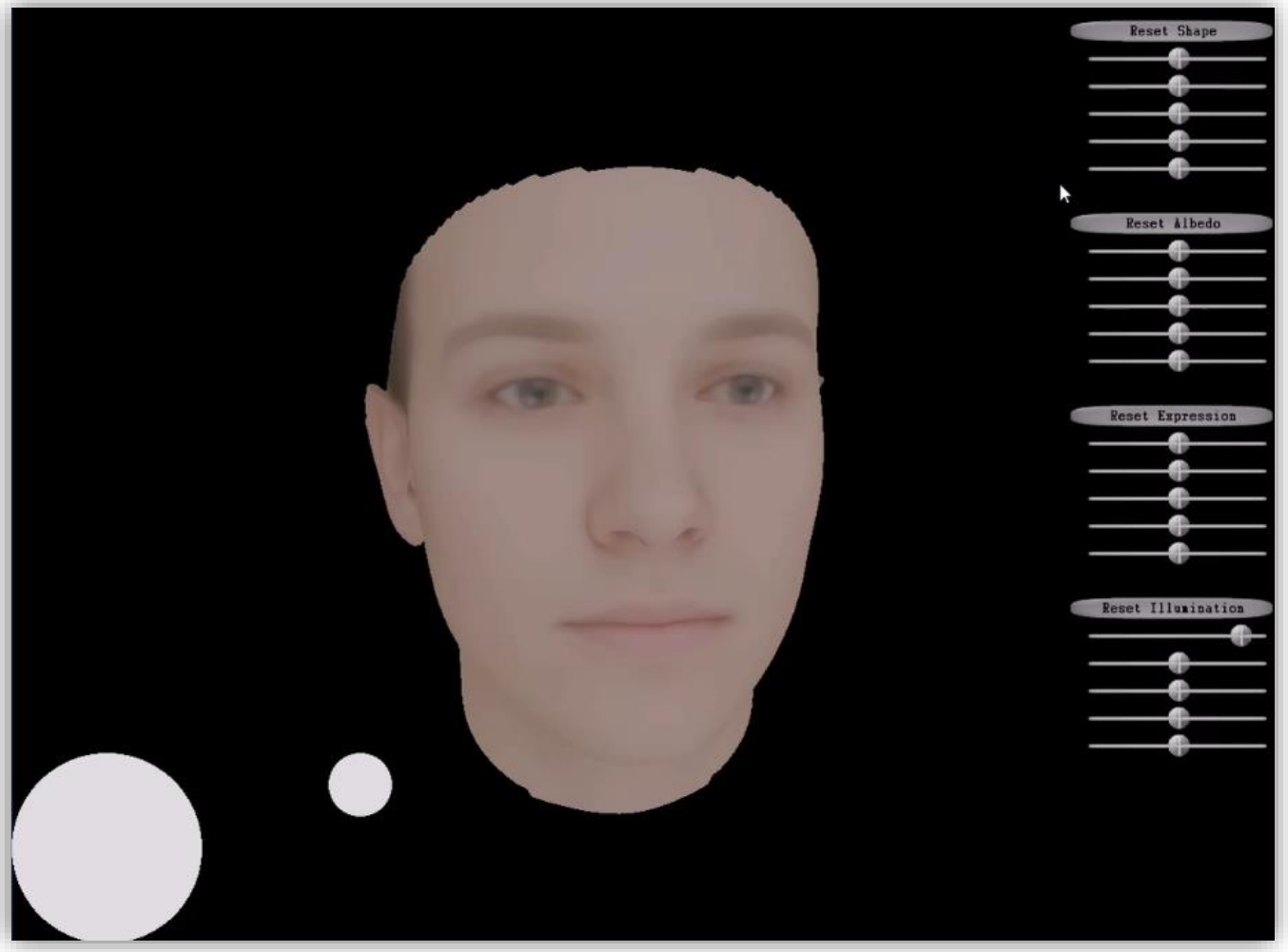
Siggraph Asia'15 [Thies et al.]: Facial Reenactment

Parametric Face Model

Shape Identity

$$\boldsymbol{P} = \begin{pmatrix} \Phi \\ \alpha \\ \\ \\ \end{pmatrix}$$

$|\boldsymbol{P}| = 6 + 80$



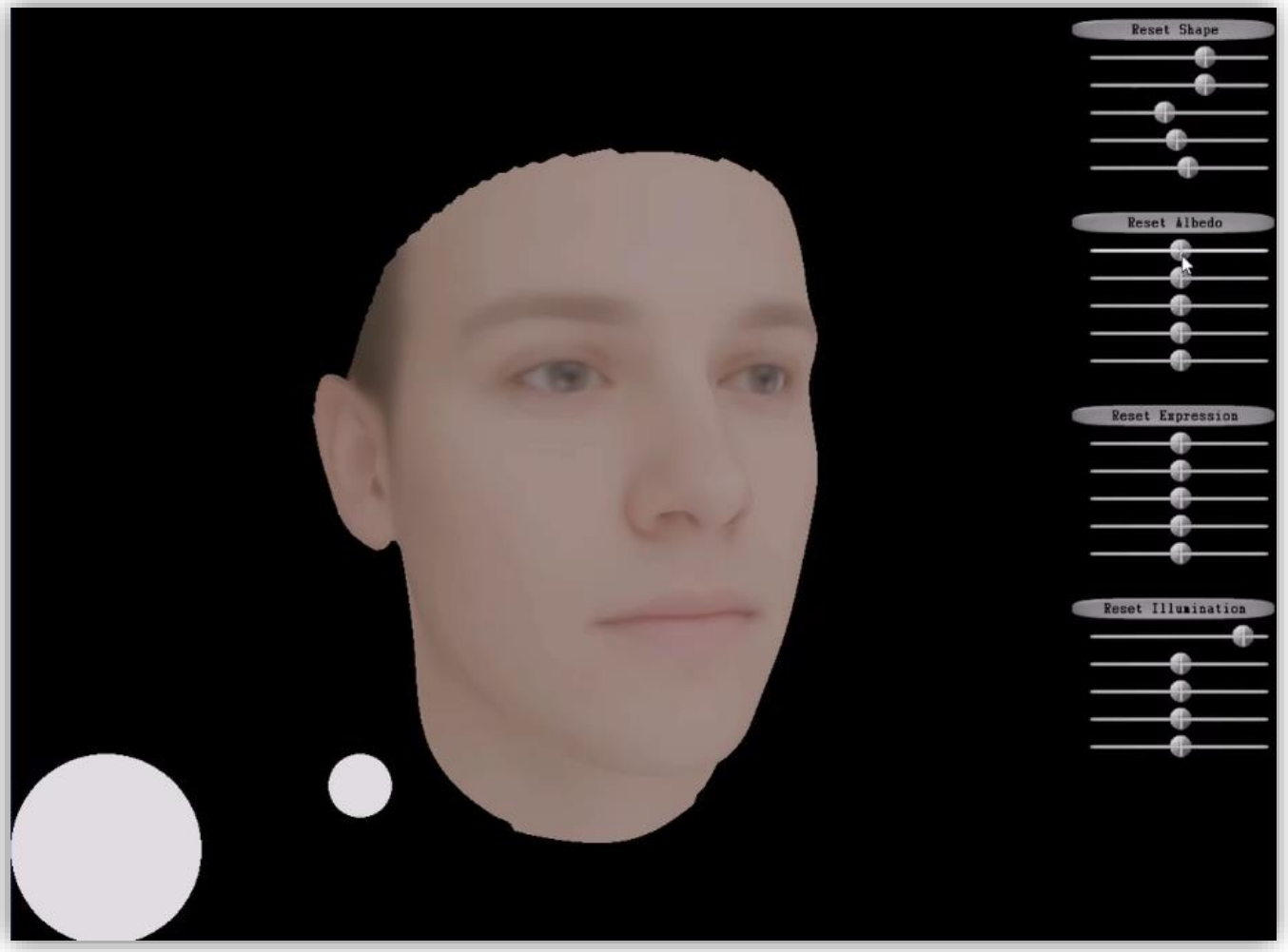
Siggraph Asia'15 [Thies et al.]: Facial Reenactment

Parametric Face Model

Material / Reflection

$$\boldsymbol{P} = \begin{pmatrix} \Phi \\ \alpha \\ \beta \end{pmatrix}$$

$|\boldsymbol{P}| = 6 + 80 + 80$



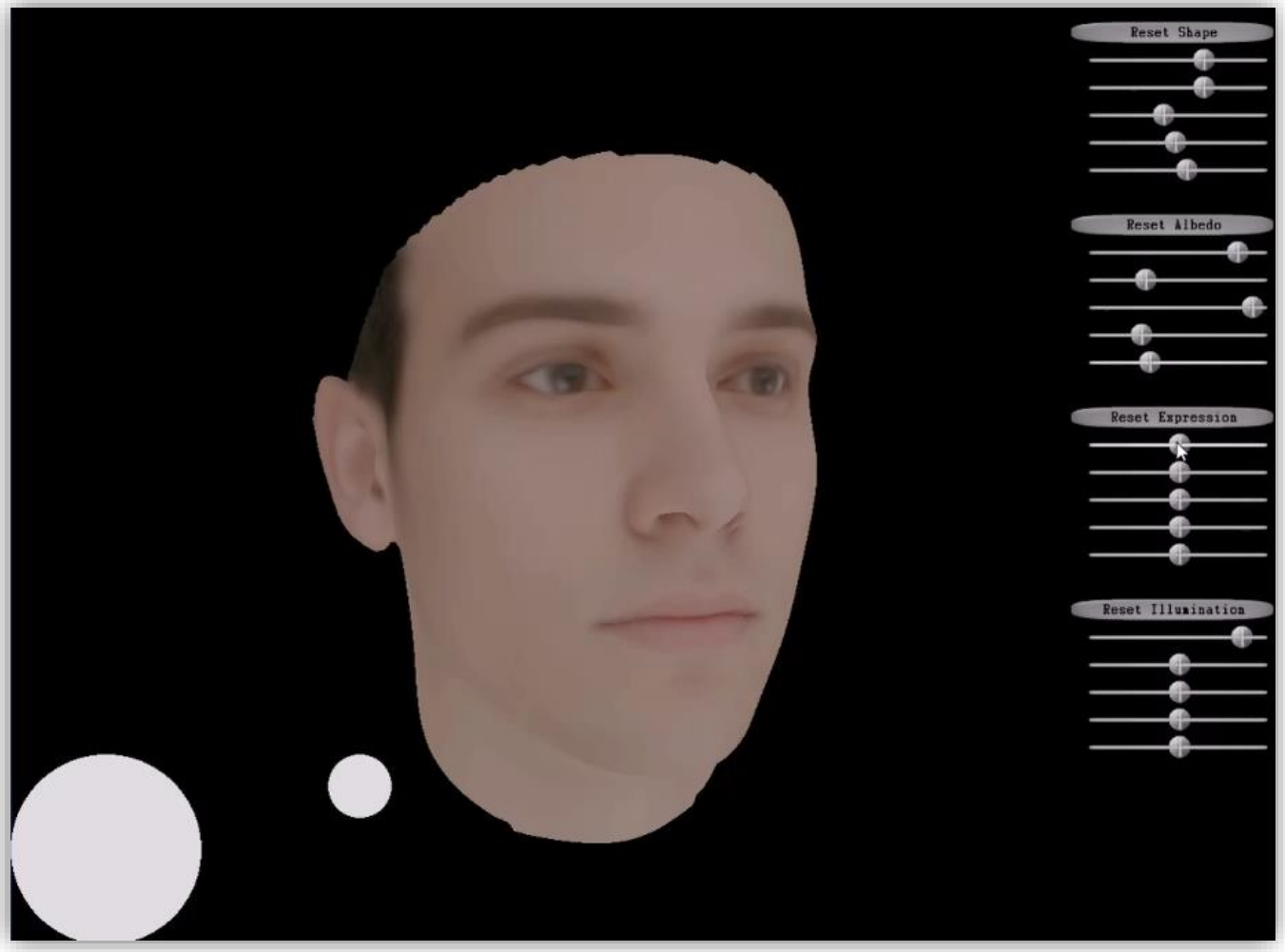
Siggraph Asia'15 [Thies et al.]: Facial Reenactment

Parametric Face Model

Expression Parameters

$$\boldsymbol{P} = \begin{pmatrix} \Phi \\ \alpha \\ \beta \\ \delta \end{pmatrix}$$

$$|\boldsymbol{P}| = 6 + 80 + 80 + 76$$



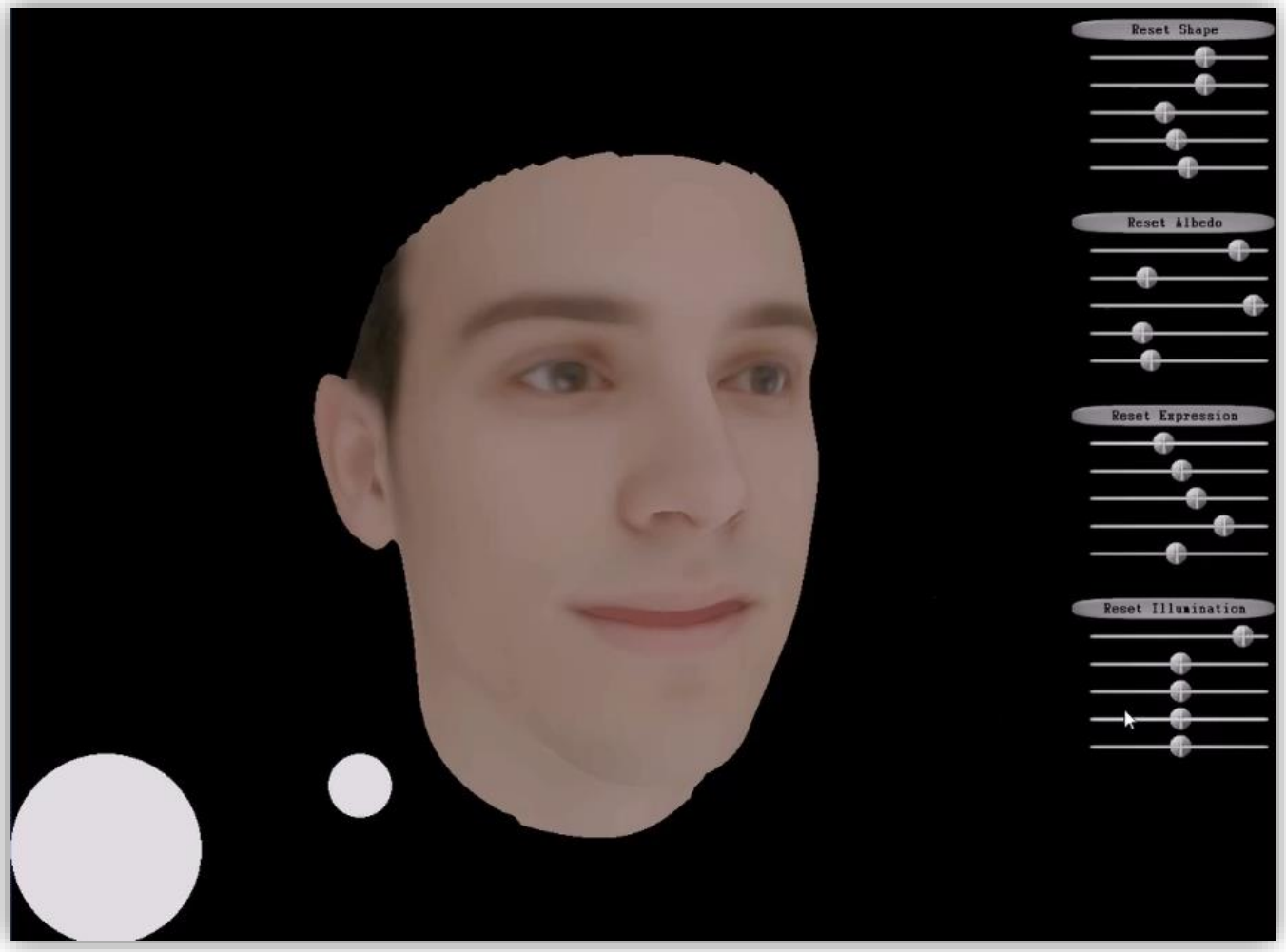
Siggraph Asia'15 [Thies et al.]: Facial Reenactment

Parametric Face Model

Lighting Parameters

$$\boldsymbol{P} = \begin{pmatrix} \Phi \\ \alpha \\ \beta \\ \delta \\ \gamma \end{pmatrix}$$

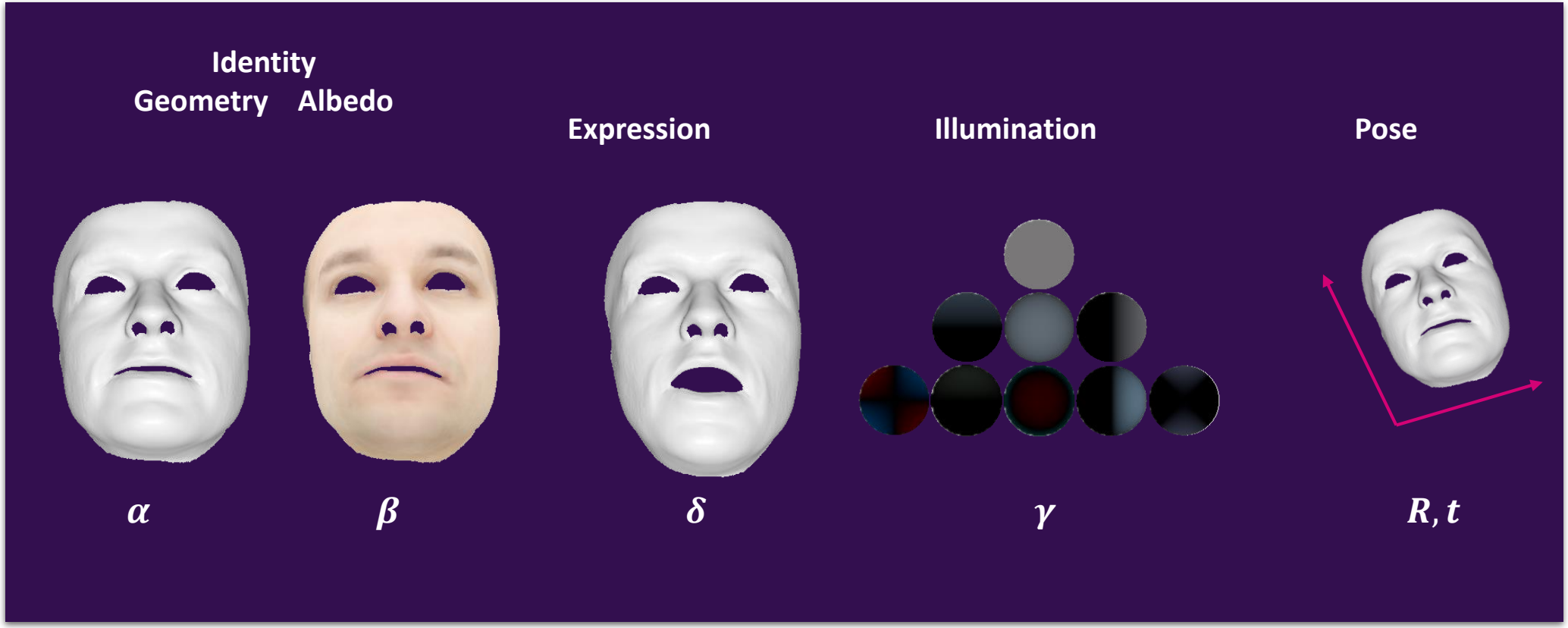
$$|\boldsymbol{P}| = 6 + 80 + 80 + 76 + 27$$



Siggraph Asia'15 [Thies et al.]: Facial Reenactment

Parametric Face Model

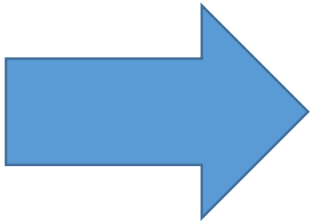
$P =$



Siggraph Asia'15 [Thies et al.]: Facial Reenactment

Parametric Face Model

$$M(P)$$



CVPR'16 (Oral) [Thies et al.]: Face2Face

Fitting Parametric Model to RGB Image

$$E(P) =$$

CVPR'16 (Oral) [Thies et al.]: Face2Face

Analysis-by-Synthesis

Given: Parametric Model $M(P)$

1. Render M with parameters P_k
2. Compute diff. between rendering and target; i.e., $E(P)$
3. Update $P_k \rightarrow P_{k+1}$; e.g., using differentiable renderer
4. If (diff > thresh) GOTO 1

CVPR'16 (Oral) [Thies et al.]: Face2Face

Inverse Rendering with Analysis-by-Synthesis

$$E(P) =$$

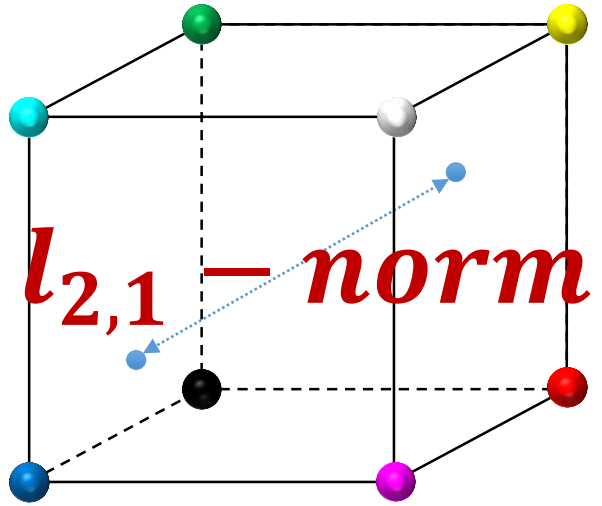
CVPR'16 (Oral) [Thies et al.]: Face2Face

Inverse Rendering with Analysis-by-Synthesis

$$E(P) = E_{col}(P)$$

Color Consistency

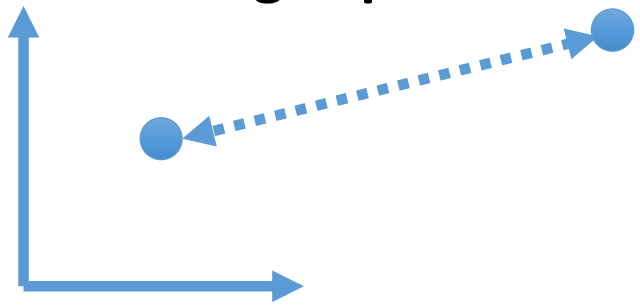
Distance in RGB Color Space



CVPR'16 (Oral) [Thies et al.]: Face2Face

Inverse Rendering with Analysis-by-Synthesis

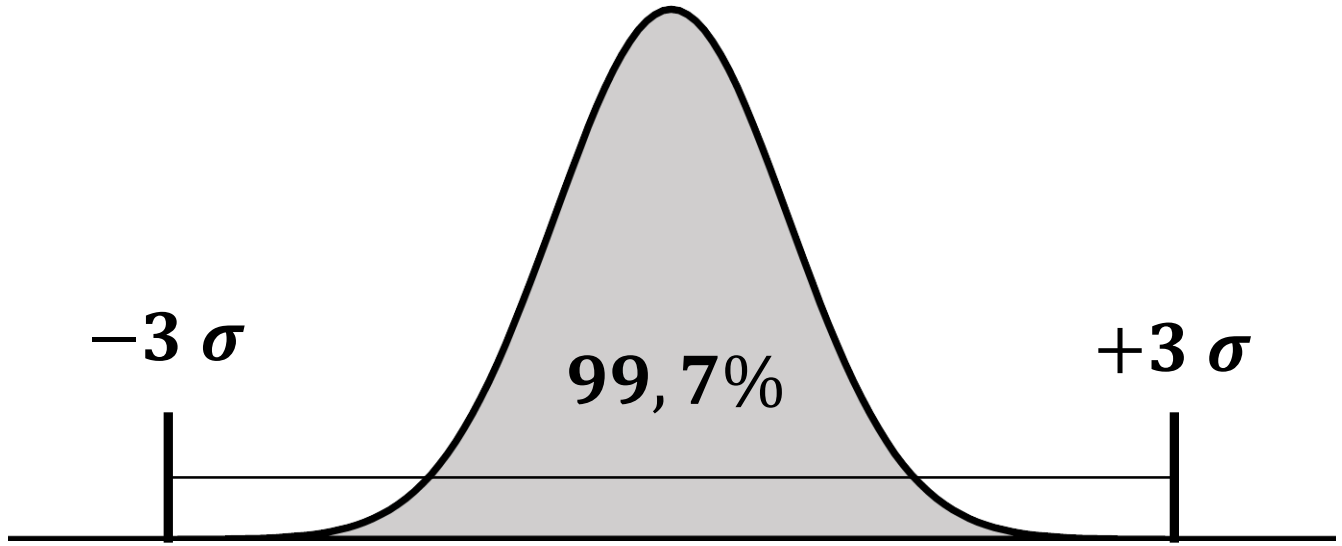
$$E(P) = \underbrace{E_{col}(P)}_{\text{Color Consistency}} + \underbrace{E_{mrk}(P)}_{\text{Feature Similarity}}$$



CVPR'16 (Oral) [Thies et al.]: Face2Face

Inverse Rendering with Analysis-by-Synthesis

$$E(P) = \underbrace{E_{col}(P)}_{\text{Color Consistency}} + \underbrace{E_{mrk}(P)}_{\text{Feature Similarity}} + \underbrace{E_{reg}(P)}_{\text{Regularization}}$$



CVPR'16 (Oral) [Thies et al.]: Face2Face

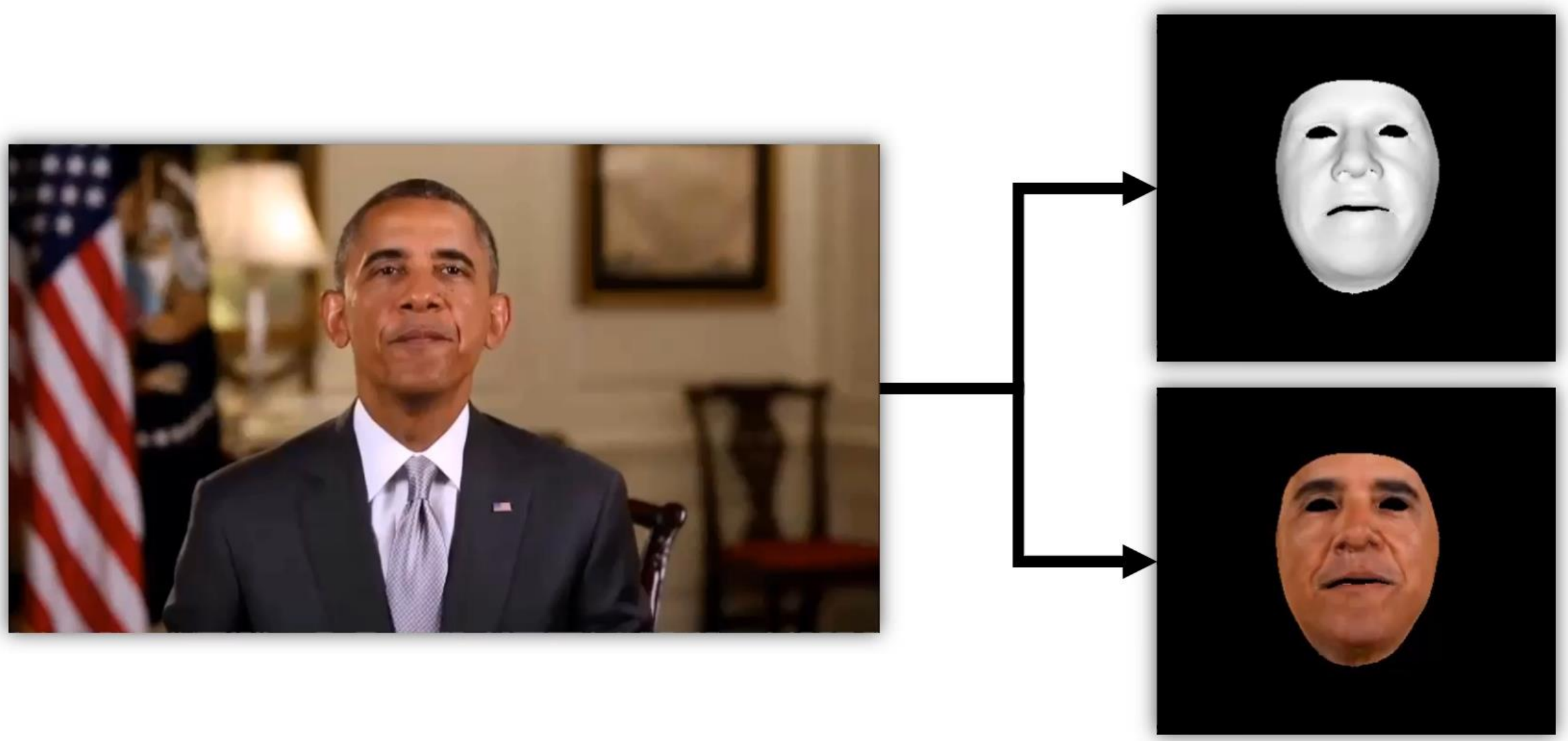
Inverse Rendering with Analysis-by-Synthesis

$$E(P) = \underbrace{E_{col}(P)}_{\text{Color Consistency}} + \underbrace{E_{mrk}(P)}_{\text{Feature Similarity}} + \underbrace{E_{reg}(P)}_{\text{Regularization}}$$

- Coarse-to-fine Gauss-Newton optimization (IRLS)
- Gradients through differentiable rendering

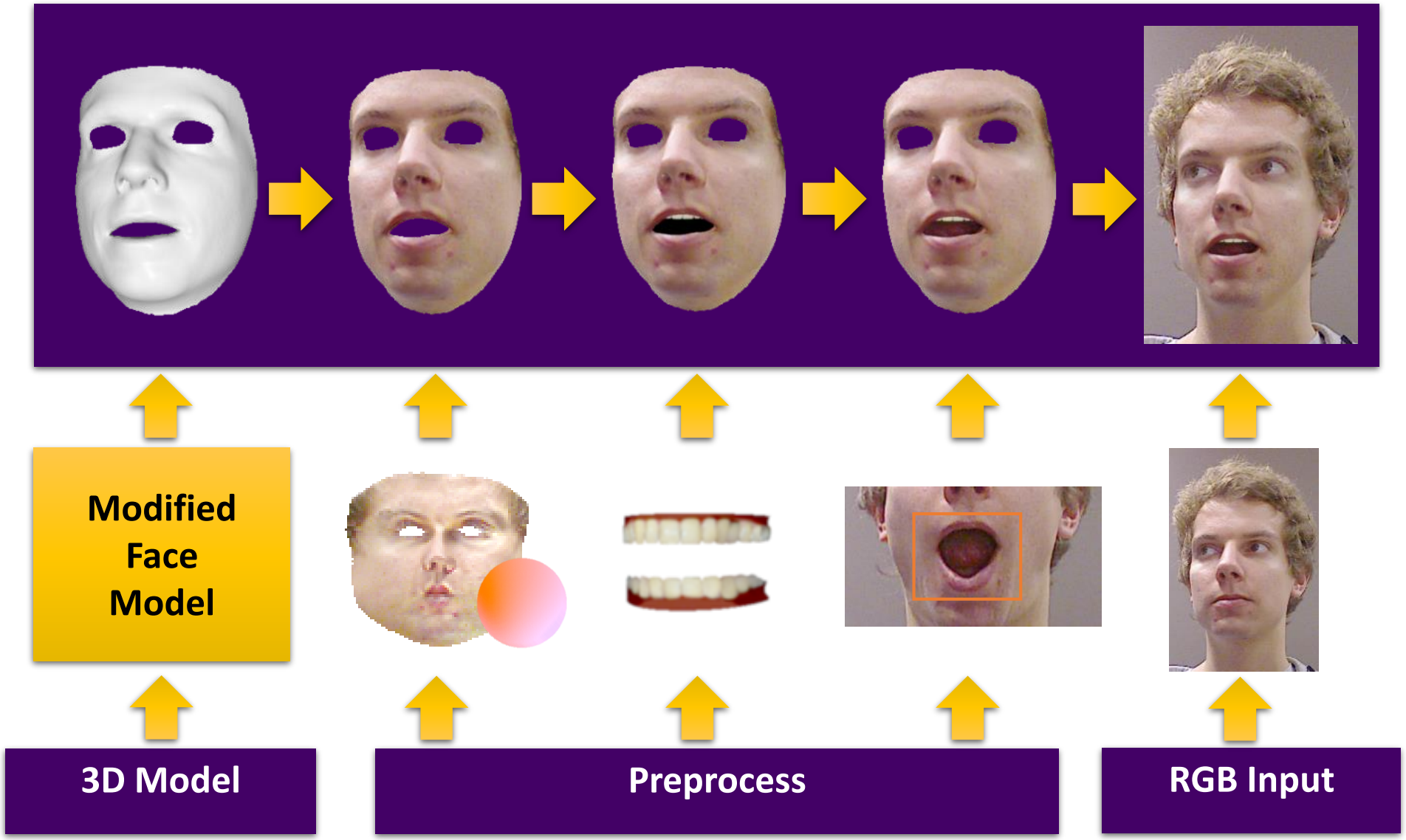
CVPR'16 (Oral) [Thies et al.]: Face2Face

Inverse Rendering with Analysis-by-Synthesis



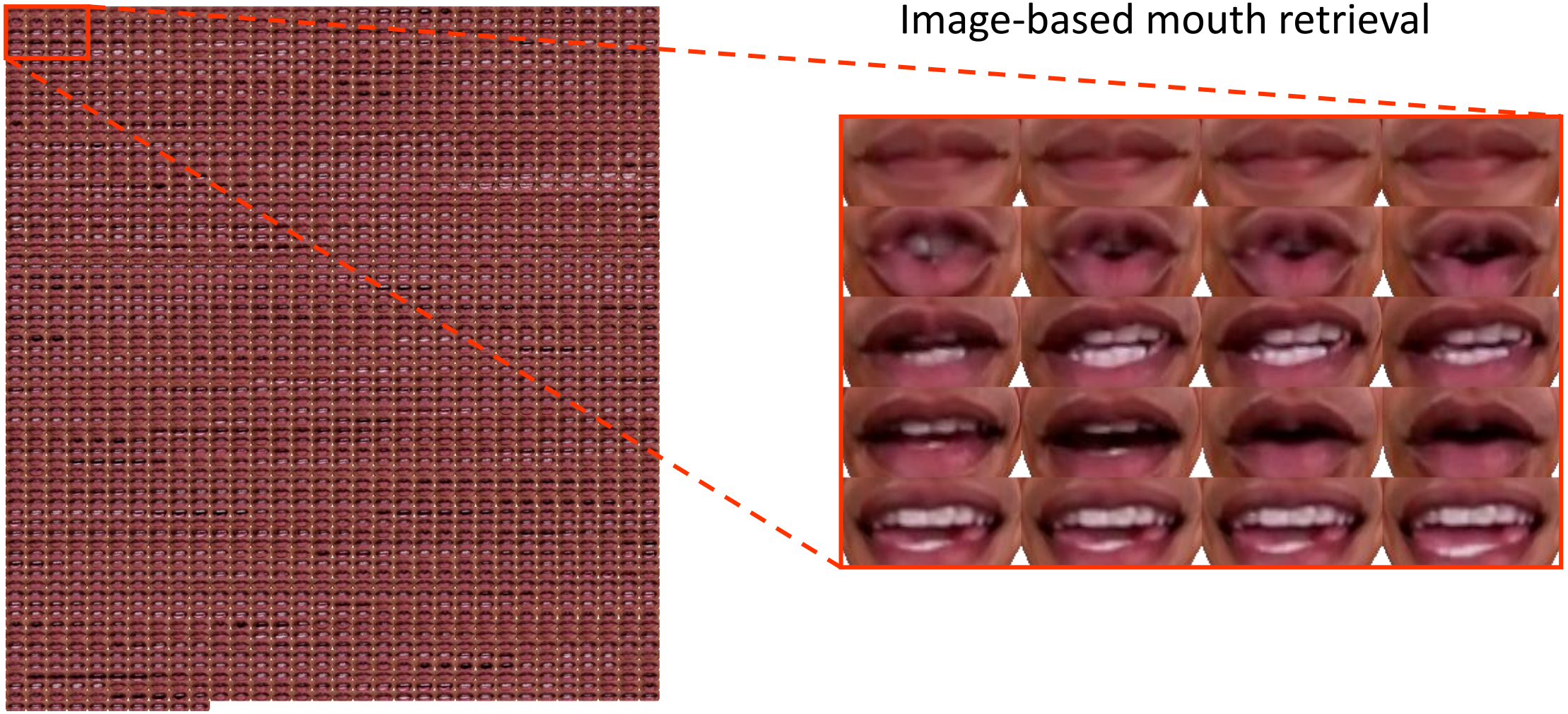
CVPR’16 (Oral) [Thies et al.]: Face2Face

3D Model + Image-based Rendering



Siggraph Asia'15 [Thies et al.]: Facial Reenactment

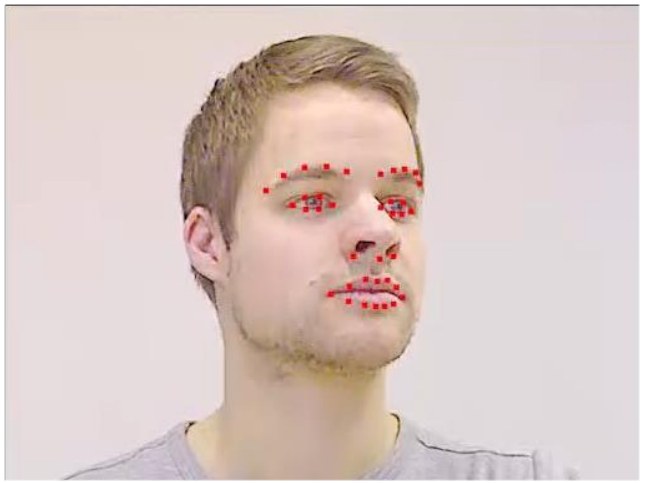
3D Model + Image-based Rendering

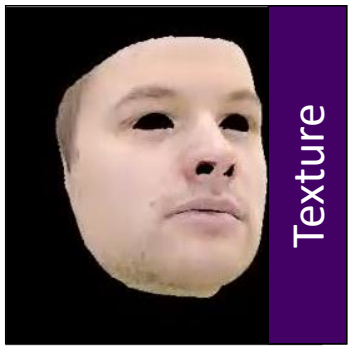


CVPR'16 (Oral) [Thies et al.]: Face2Face

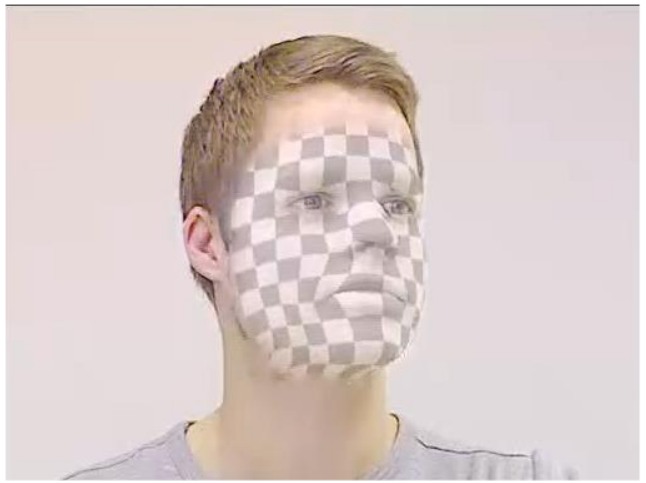
3D Model + Image-based Rendering

Input



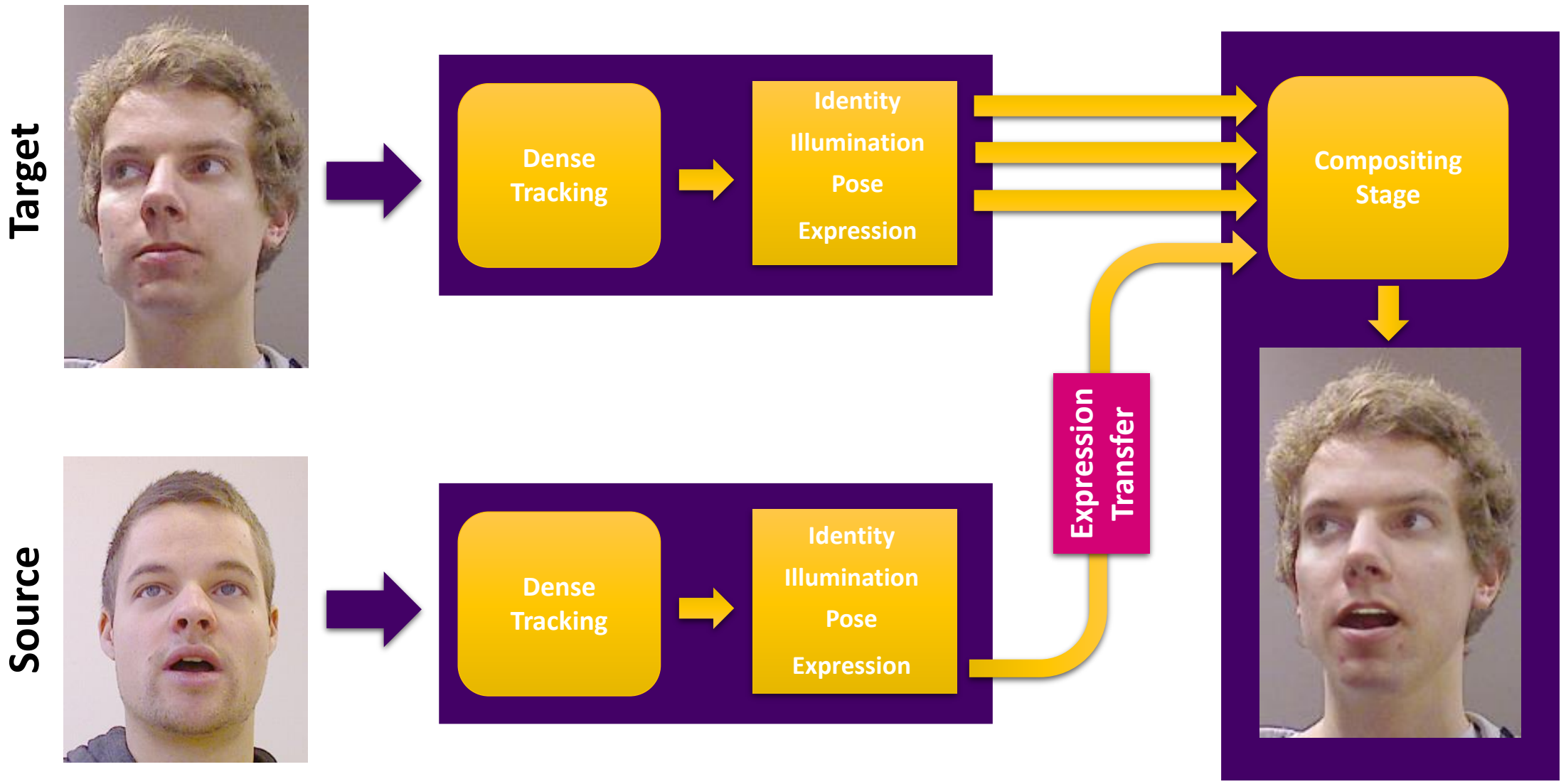


Output



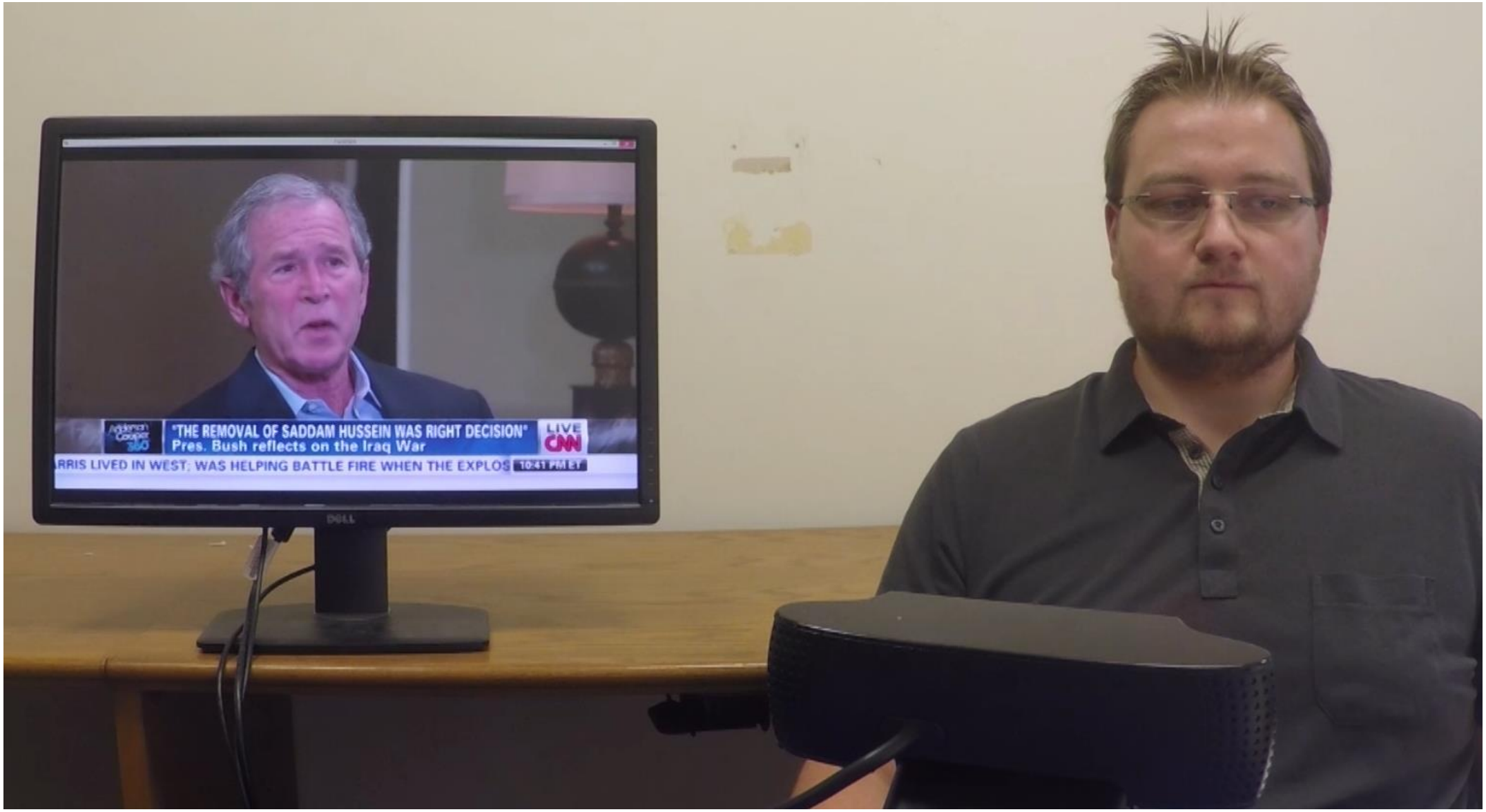
Siggraph Asia'15 [Thies et al.]: Facial Reenactment

Facial Expression Transfer



Siggraph Asia'15 [Thies et al.]: Facial Reenactment

Face2Face



CVPR'16 (Oral) [Thies et al.]: Face2Face

HeadOn: Reenactment of Portrait Videos

Source Actor

Reenacted Proxy

Reenacted Output

Siggraph'18 [Thies et al.]: HeadOn

HeadOn: Reenactment of Portrait Videos

Source Actor

Reenacted Proxy

Reenacted Output

Siggraph'18 [Thies et al.]: HeadOn

HeadOn: Reenactment of Portrait Videos

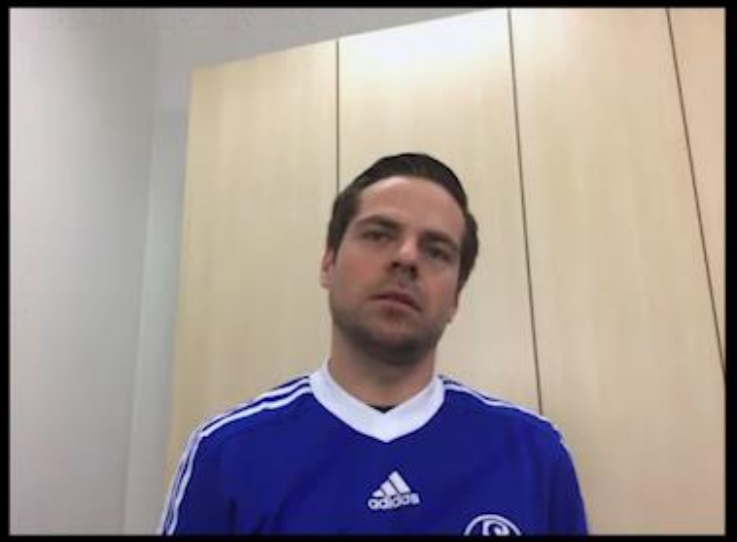
Source Actor

Reenacted Proxy

Reenacted Output

Siggraph'18 [Thies et al.]: HeadOn

HeadOn: Reenactment of Portrait Videos



Source Actor

Reenacted Proxy

Reenacted Output

Siggraph'18 [Thies et al.]: HeadOn

Analysis-by-Synthesis

Parametric model needs to be flexible

-> there needs to be a P that re-creates captured RGB input

Optimizable

-> Must be able to find good optimum in energy landscape $E(P)$

Incompleteness

-> Image-based tricks to fix 3D artifacts are unsatisfactory

Generative Neural Networks

Generative Neural Networks

Over-parameterized models

-> can re-create input

Generative Neural Networks

Over-parameterized models
-> can re-create input

Discriminator loss $J^{(D)} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log\left(1 - D\left(G(\boldsymbol{z})\right)\right)$

Generator loss $J^{(G)} = -J^{(D)}$

GANs [Goodfellow et al. 14], Pix2Pix [Isola et al. 17], ProGAN [Karras et al. 18], …

Generative Neural Networks

Over-parameterized models
-> can re-create input

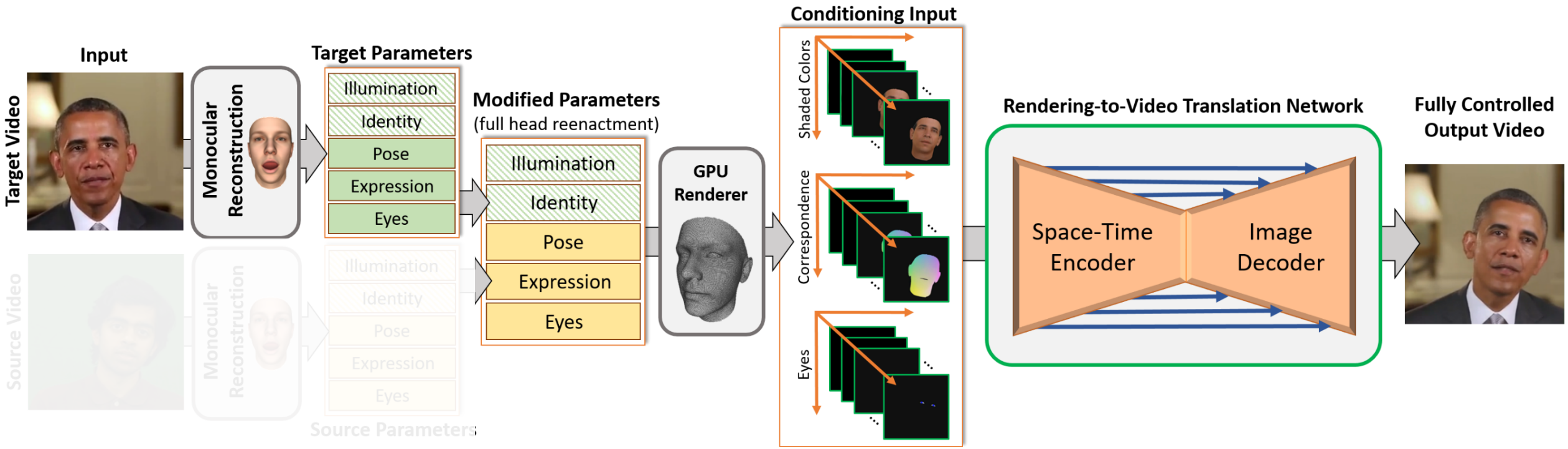
No explicit no control
-> struggle with videos

Discriminator loss $J^{(D)} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log\left(1 - D\left(G(\boldsymbol{z})\right)\right)$

Generator loss $J^{(G)} = -J^{(D)}$

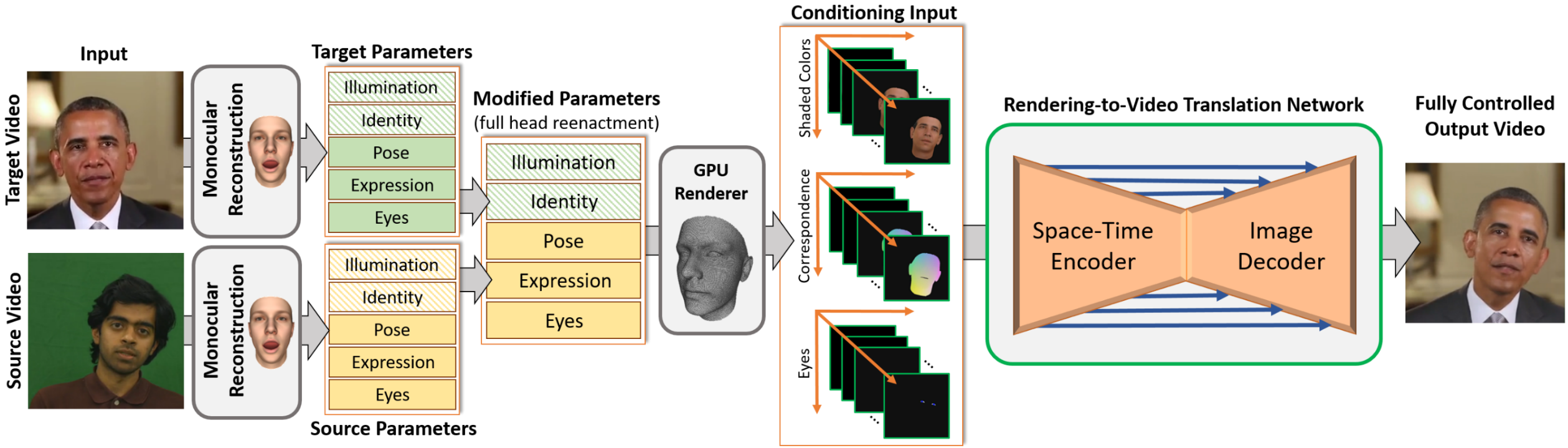
GANs [Goodfellow et al. 14], Pix2Pix [Isola et al. 17], ProGAN [Karras et al. 18], …

Conditional GANs



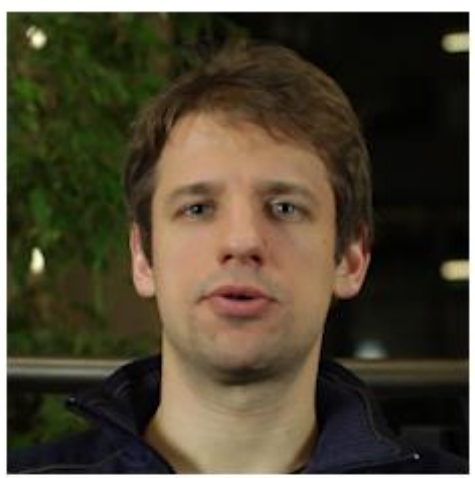
Siggraph'18 [Kim et al.]: Deep Video Portraits

Conditional GANs

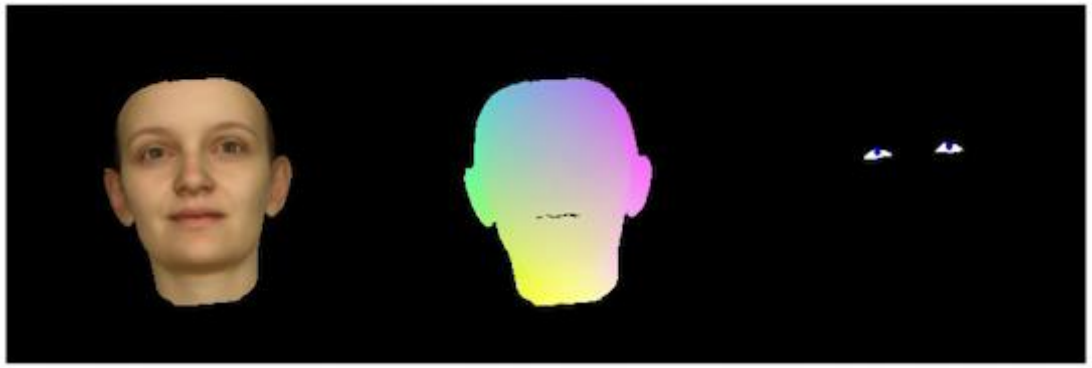


Siggraph'18 [Kim et al.]: Deep Video Portraits

Conditioning on Face Reconstruction



Source Sequence



Conditioning Images

Result

Neural Network converts synthetic data to realistic video

Siggraph'18 [Kim et al.]: Deep Video Portraits

Conditioning on Face Reconstruction

Source Sequence

Conditioning Images

Result

Siggraph'18 [Kim et al.]: Deep Video Portraits

Conditioning on Face Reconstruction

Source Sequence

Conditioning Images

Result

Siggraph'18 [Kim et al.]: Deep Video Portraits

Video Editing

Siggraph’18 [Kim et al.]: Deep Video Portraits

Other cGANs

Other cGANs

- cGANs work with different input

- Requires consistent input
  i.e., accurate tracking

Other cGANs

- cGANs work with different input

- Requires consistent input
  i.e., accurate tracking

[Chan et al. 18] Everybody Dance Now

Other cGANs

- cGANs work with different input

- Requires consistent input
  i.e., accurate tracking

[Chan et al. 18] Everybody Dance Now

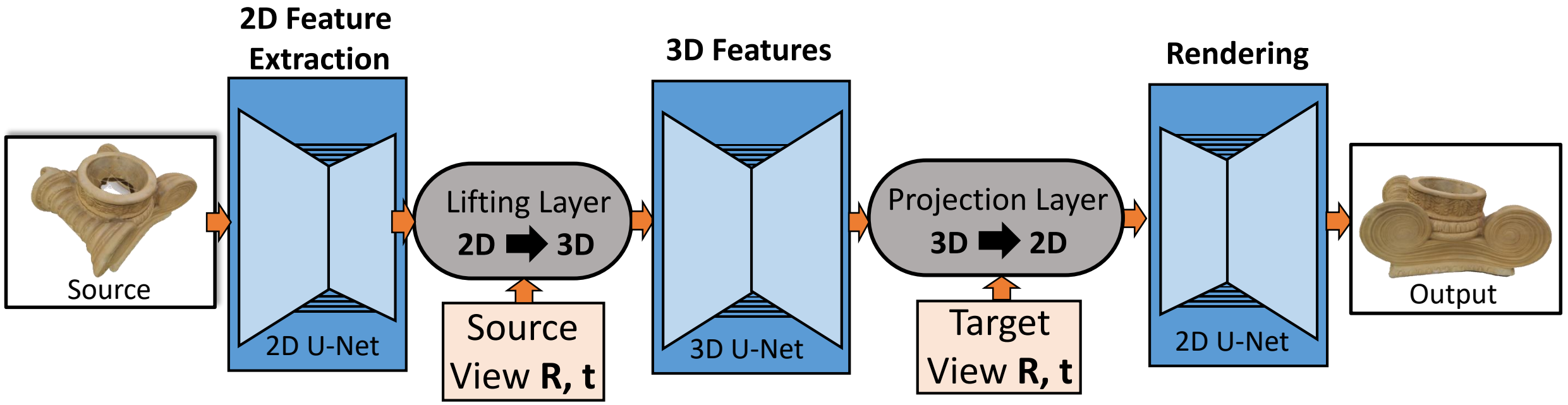
Other cGANs

- cGANs work with different input
- Requires consistent input
  i.e., accurate tracking
- Network has no explicit 3D notion

[Chan et al. 18] Everybody Dance Now

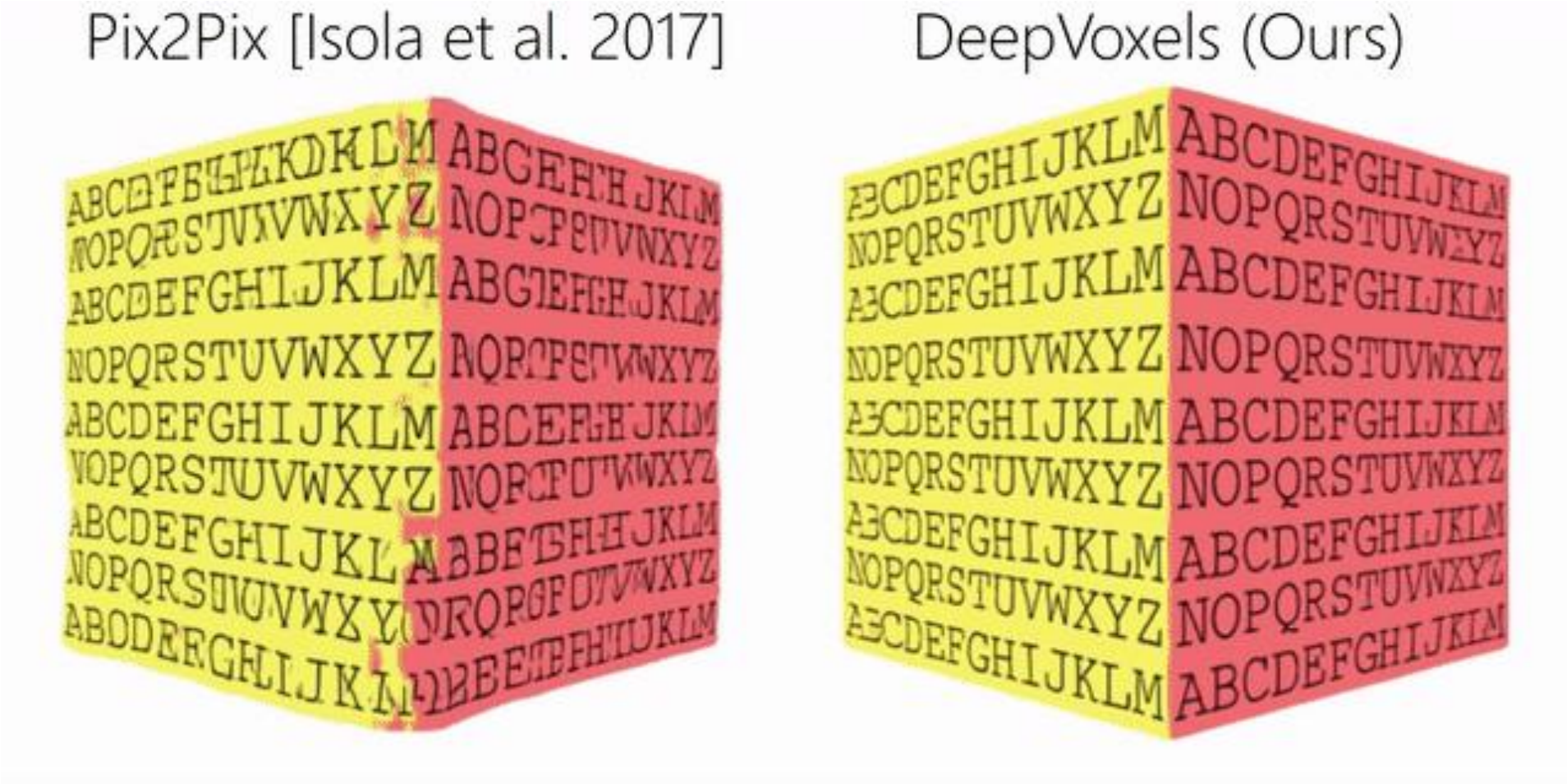
Videos still challenging for cGANs…

DeepVoxels: Explicit 3D Features



Simplified overview for novel view synthesis

DeepVoxels: Explicit 3D Features

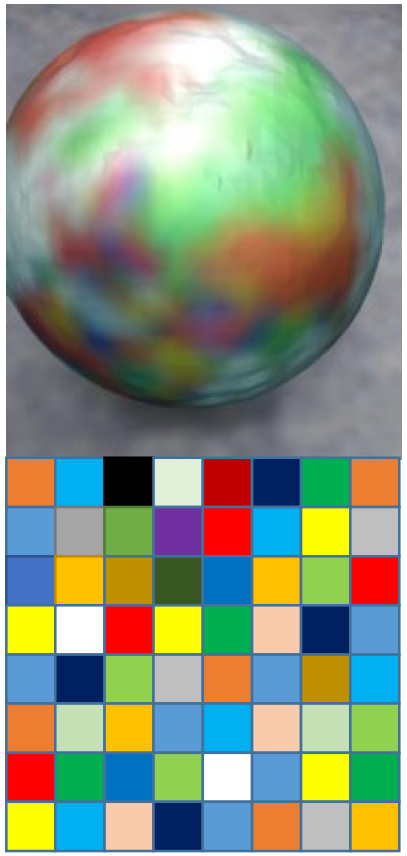


CVPR'19 (Oral) [Sitzmann et al.]: DeepVoxels

Neural Textures: Features on 3D Mesh

Neural Textures: Features on 3D Mesh

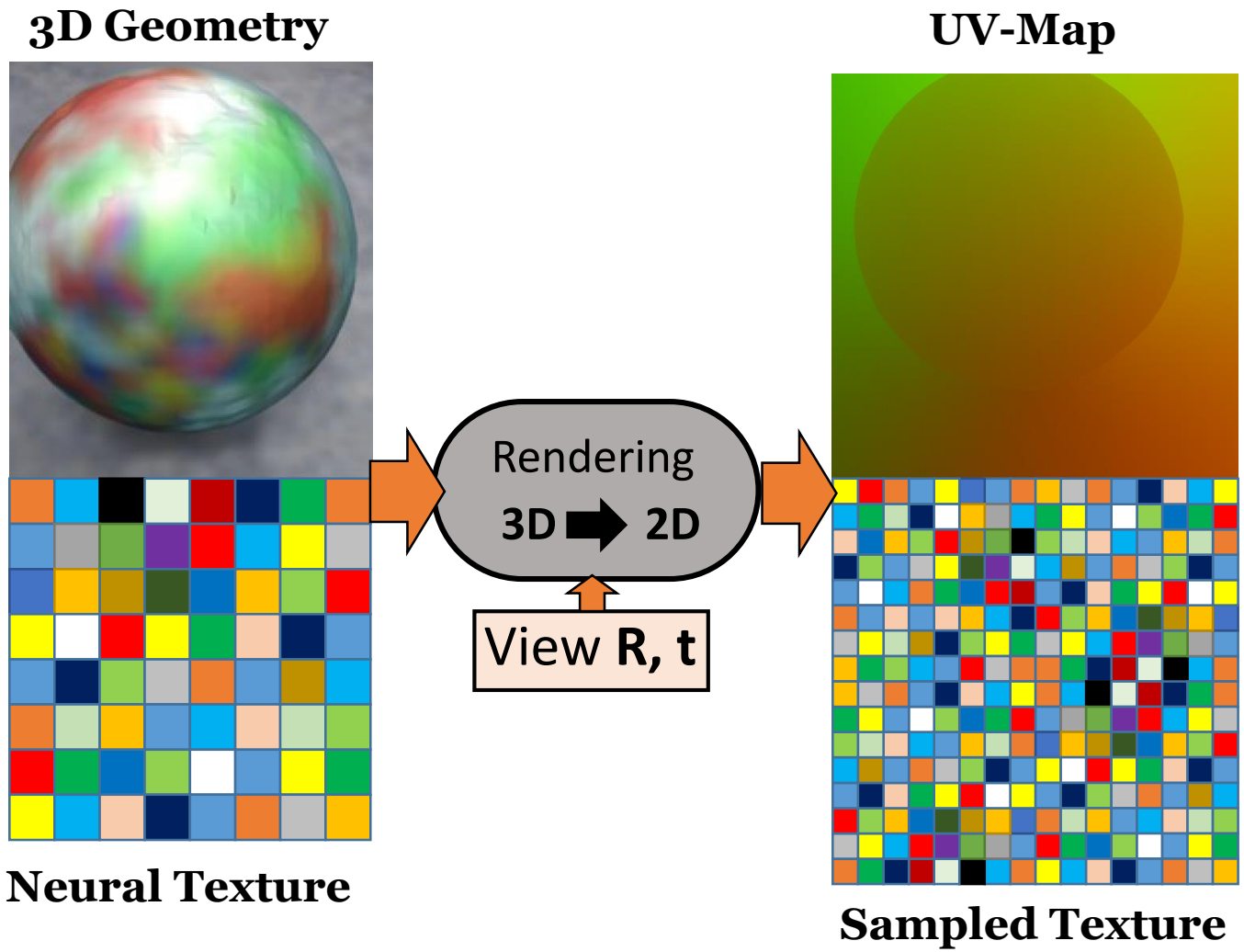
3D Geometry



Neural Texture

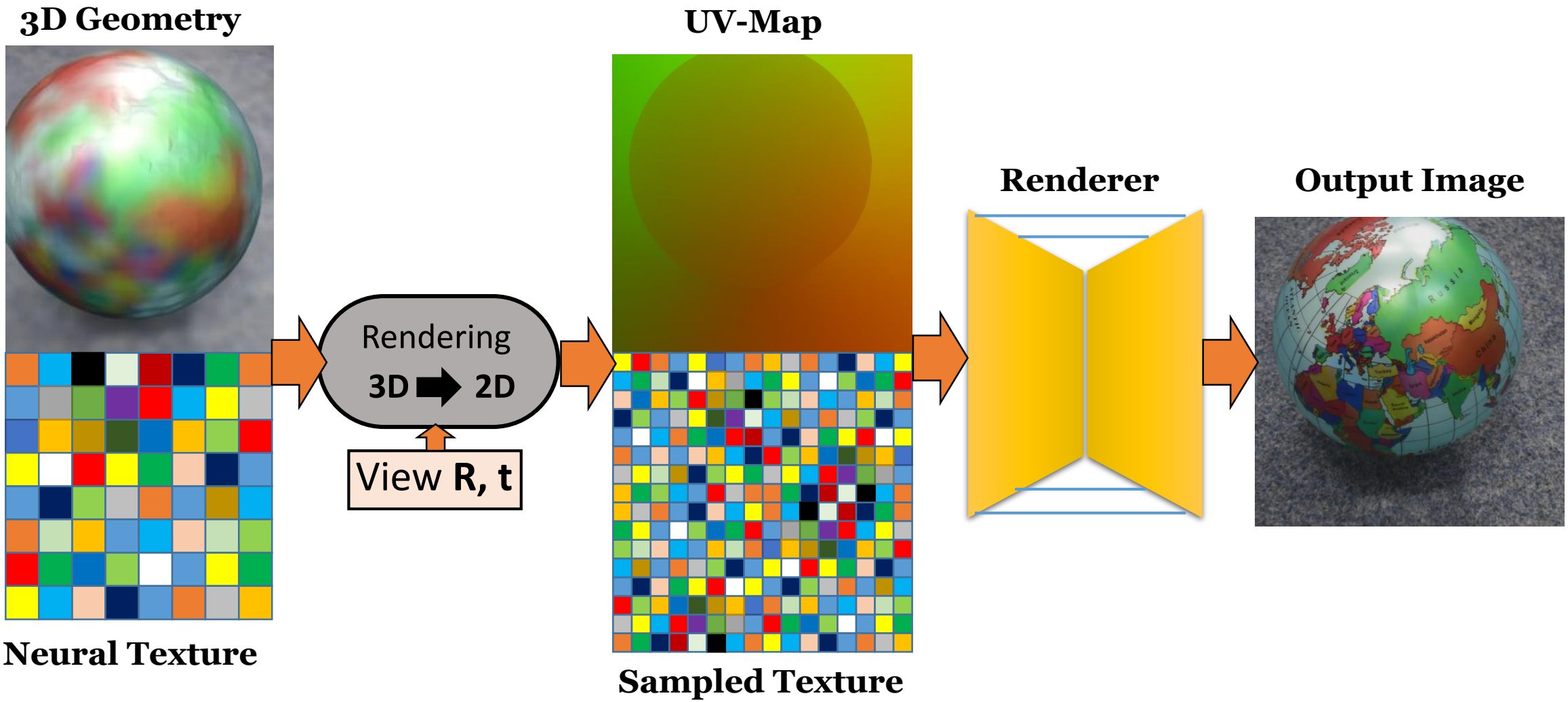
Siggraph'19 [Thies et al.]: Neural Textures

Neural Textures: Features on 3D Mesh



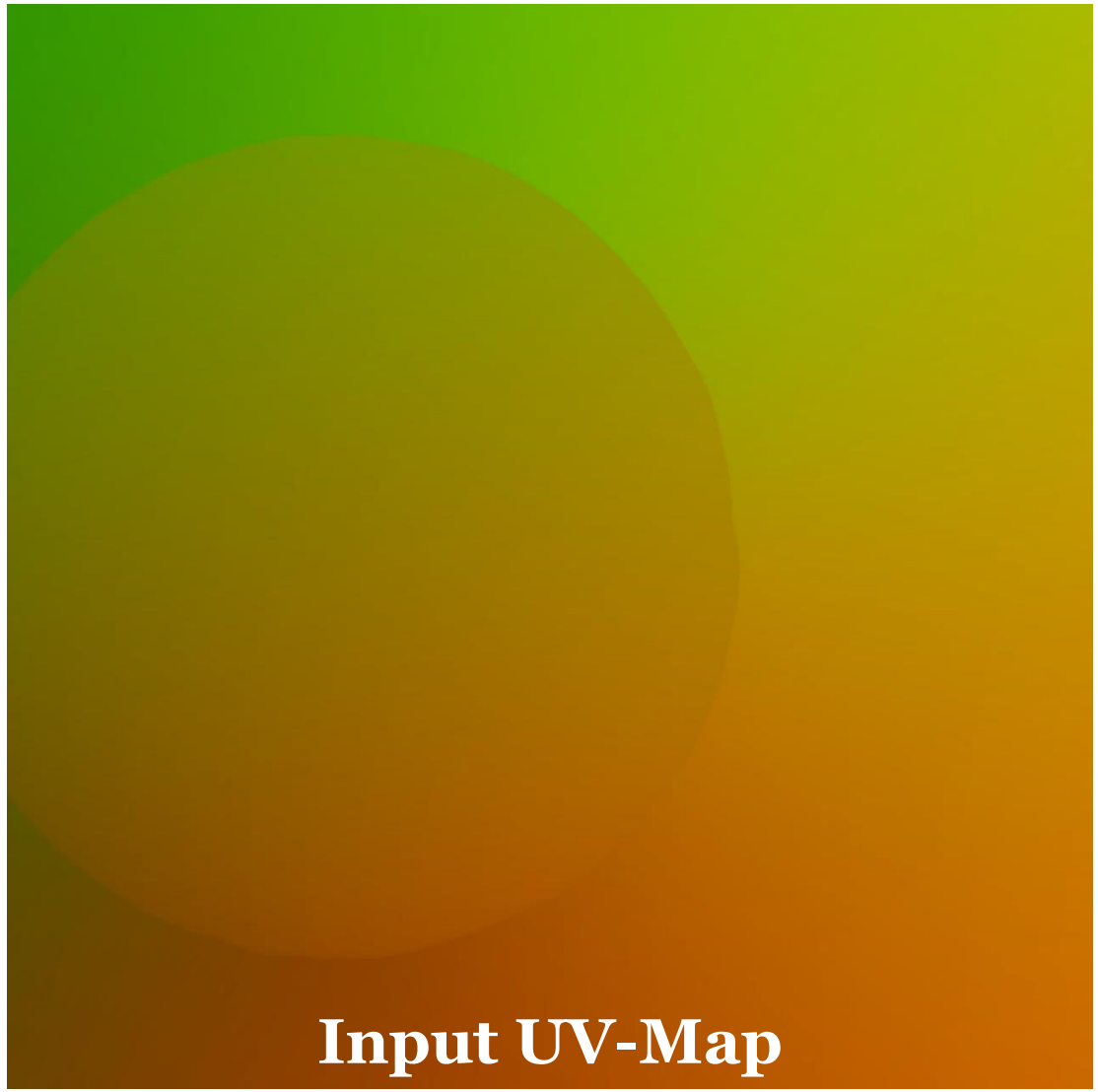
Siggraph'19 [Thies et al.]: Neural Textures

Neural Textures: Features on 3D Mesh



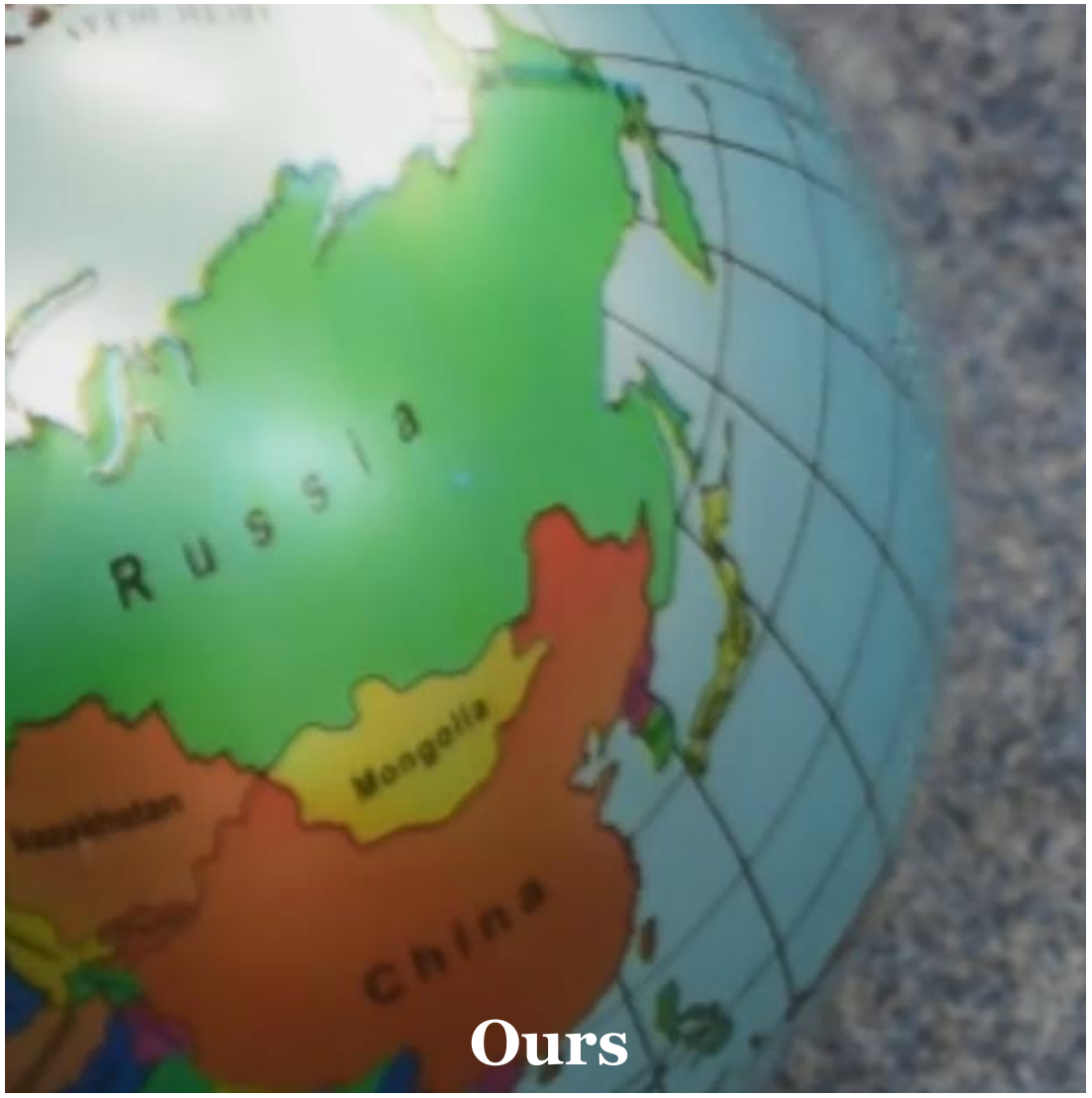
Siggraph'19 [Thies et al.]: Neural Textures

Novel View-Point Synthesis



Siggraph'19 [Thies et al.]: Neural Textures

Novel View-Point Synthesis



Siggraph'19 [Thies et al.]: Neural Textures

Scene Editing

Input Sequence

Geometry Editing

Siggraph'19 [Thies et al.]: Neural Textures

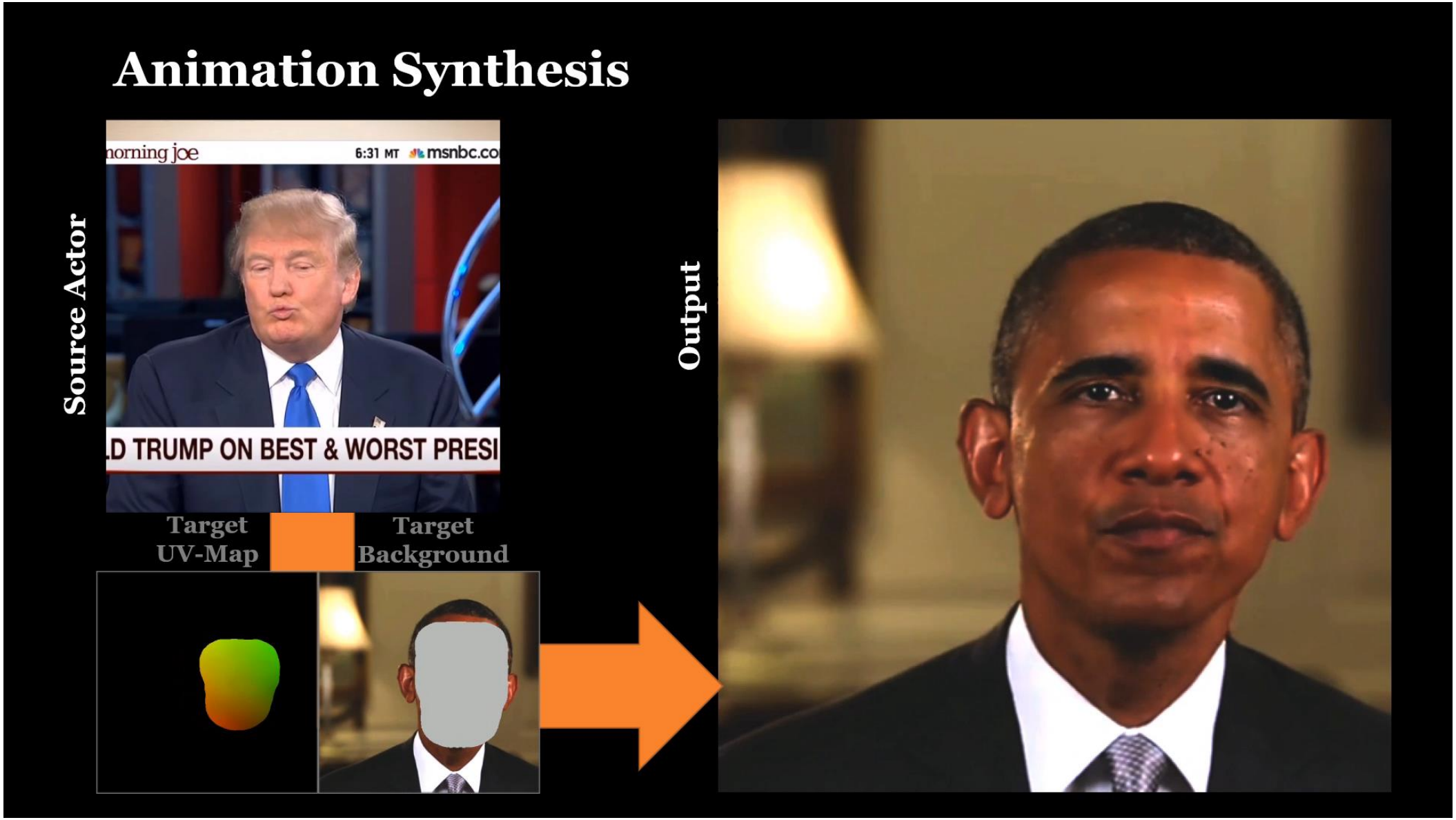
Scene Editing

Siggraph'19 [Thies et al.]: Neural Textures

Scene Editing

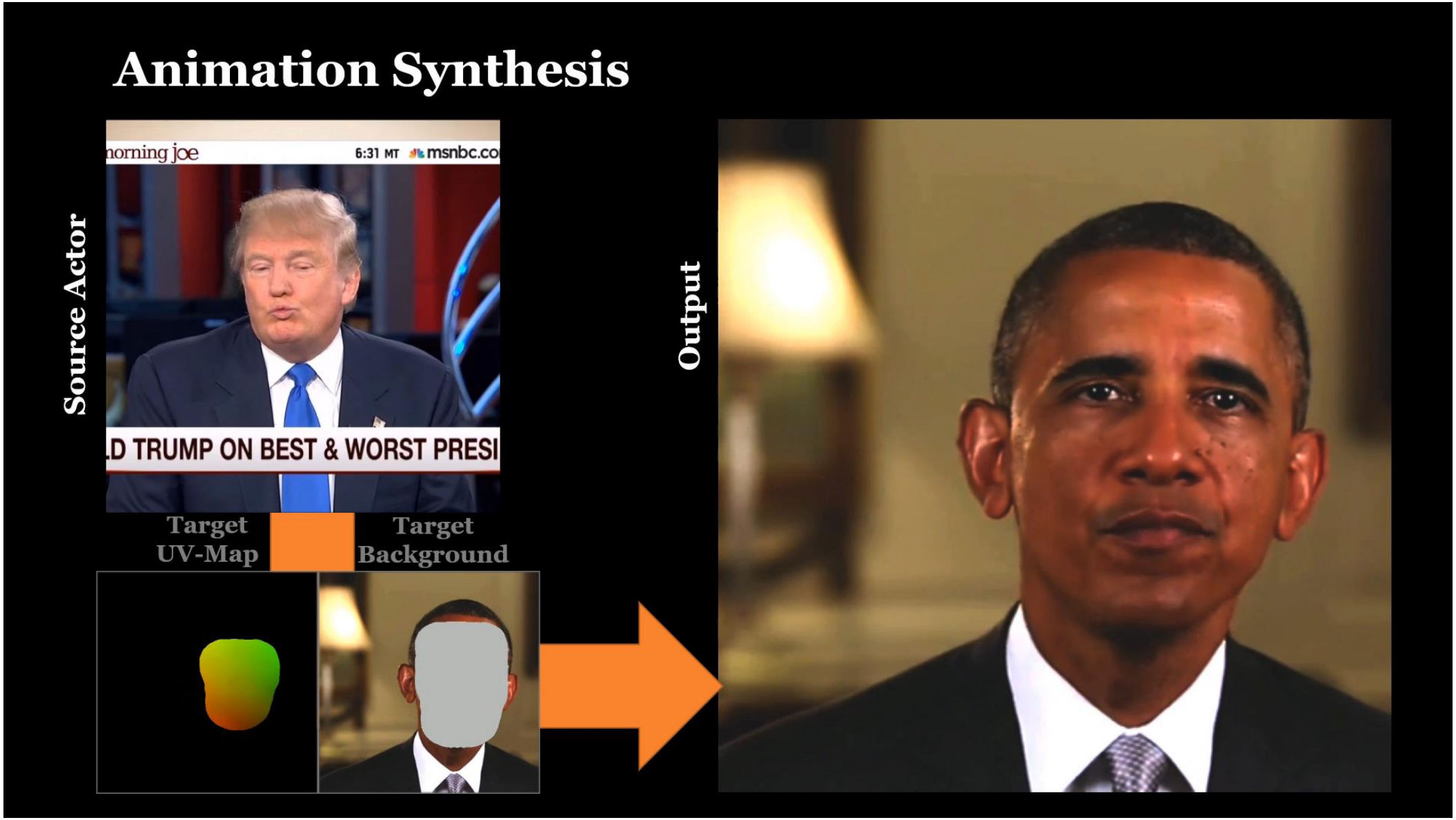
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



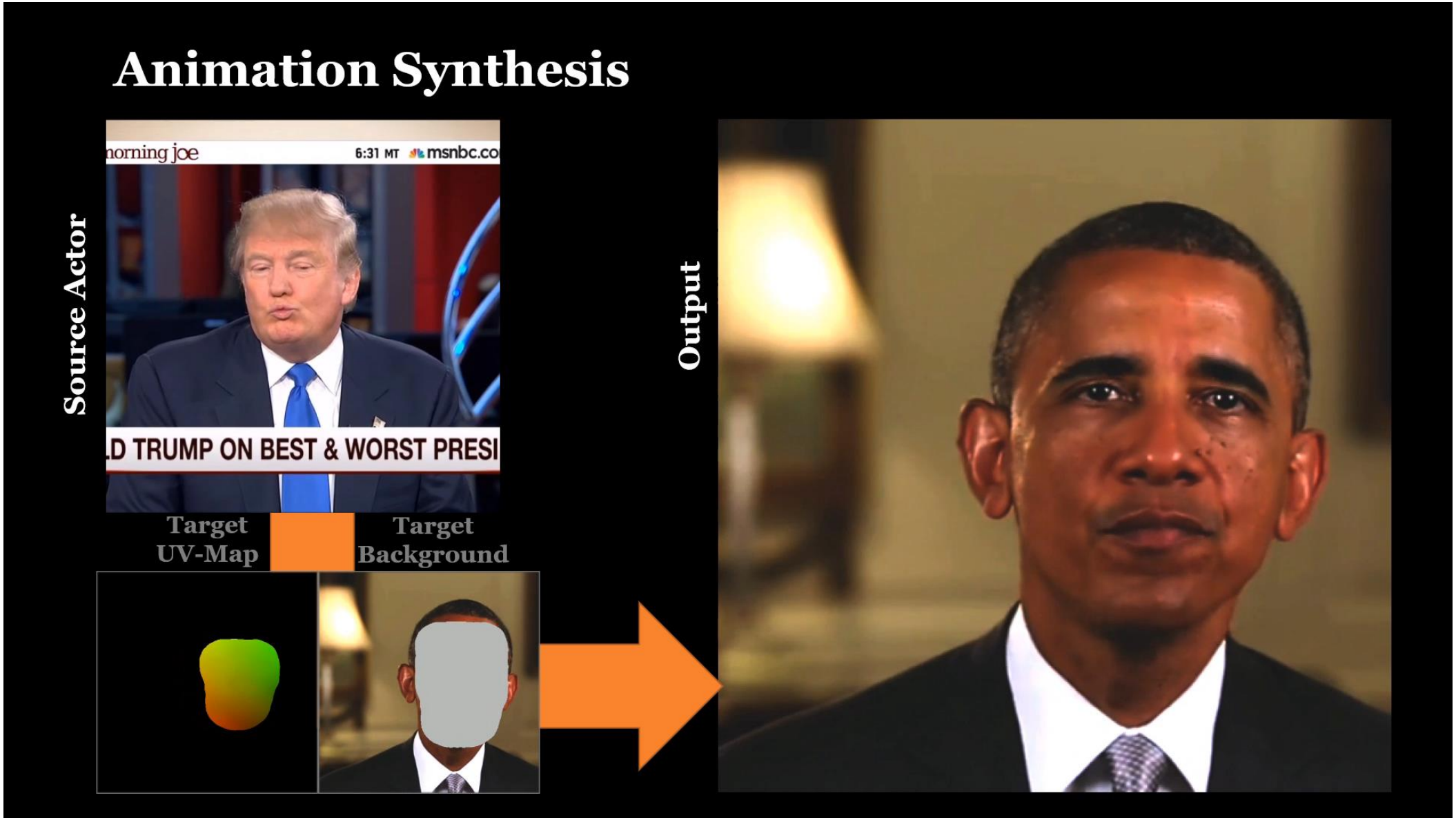
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



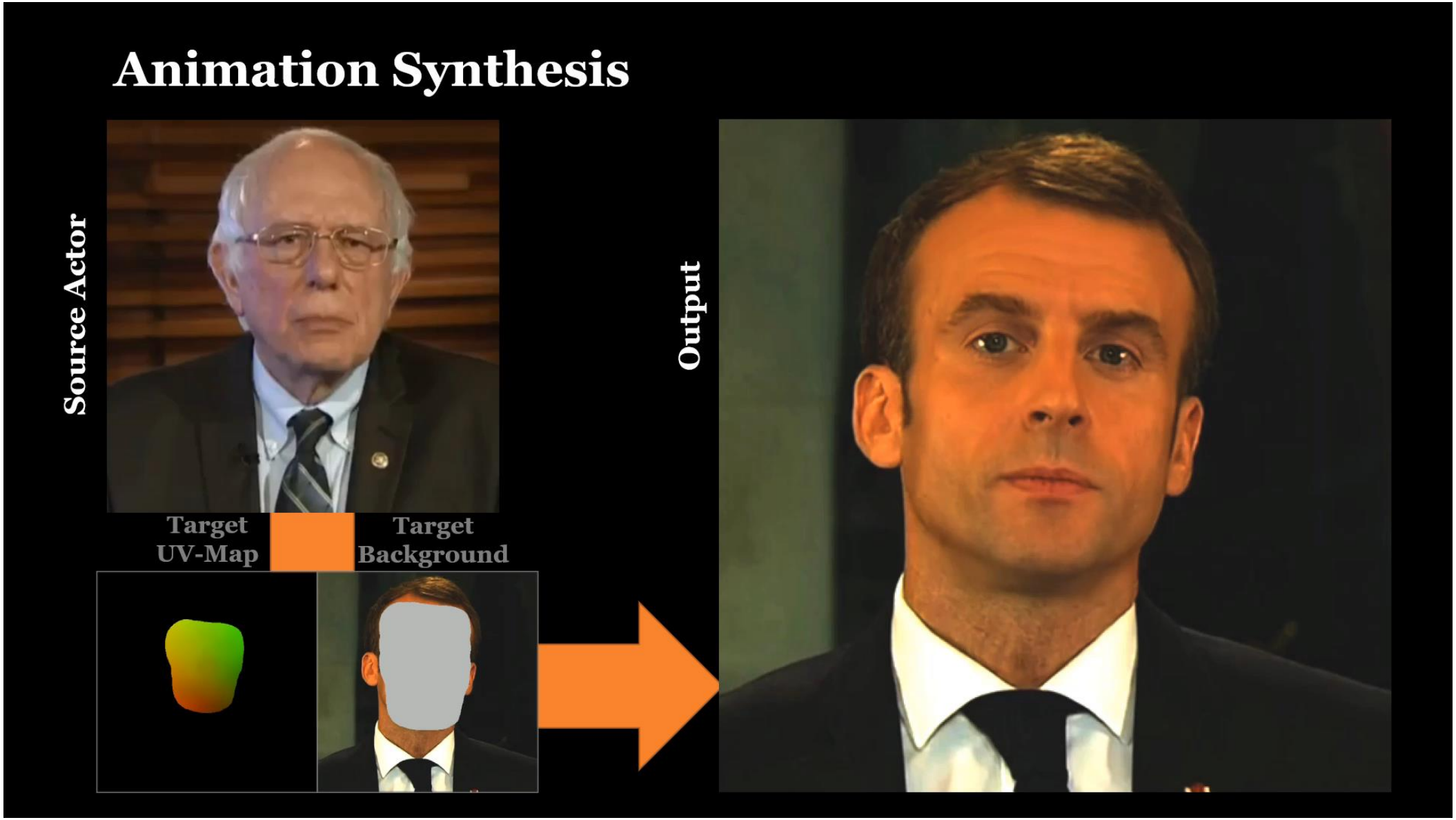
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



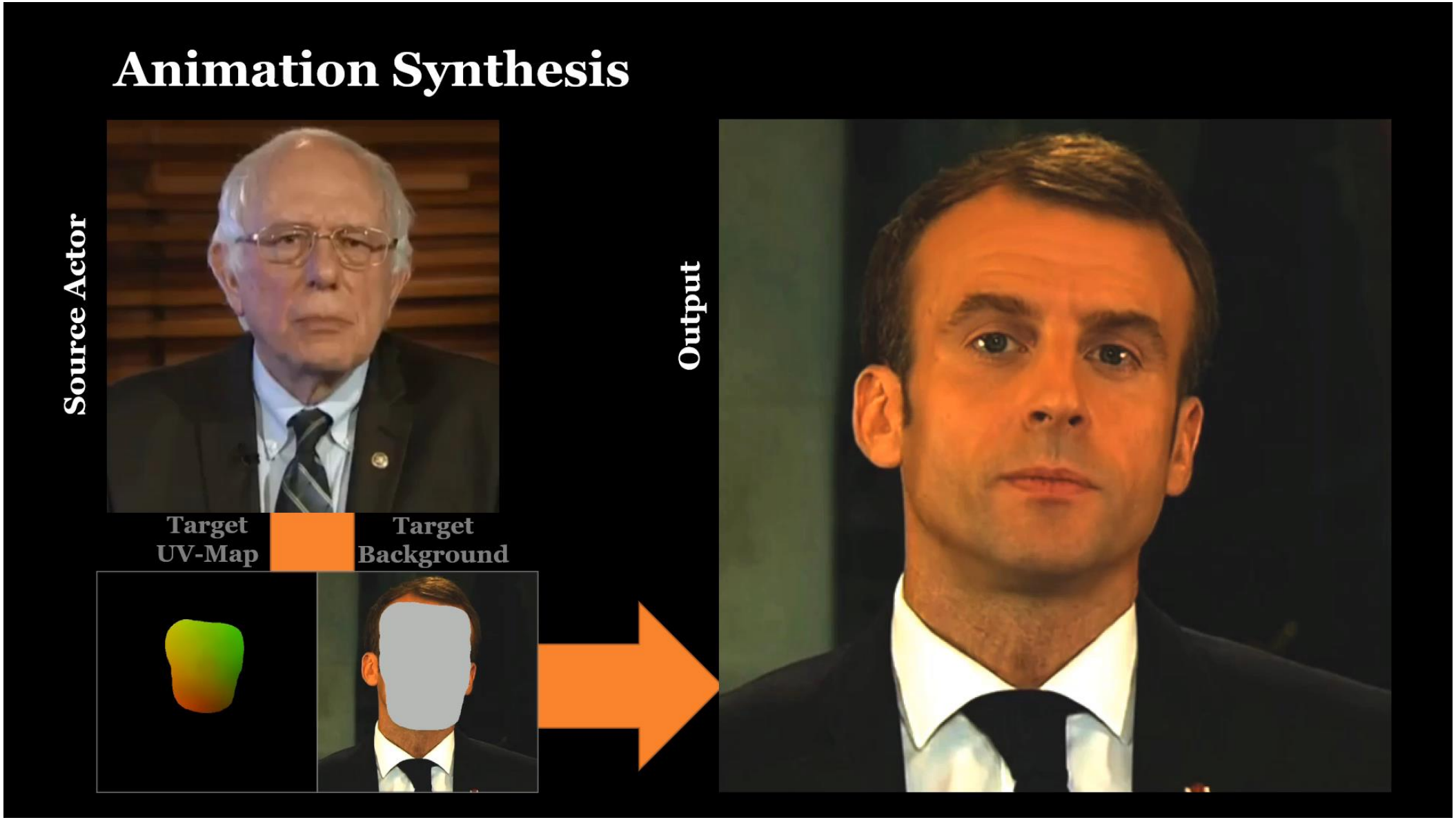
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



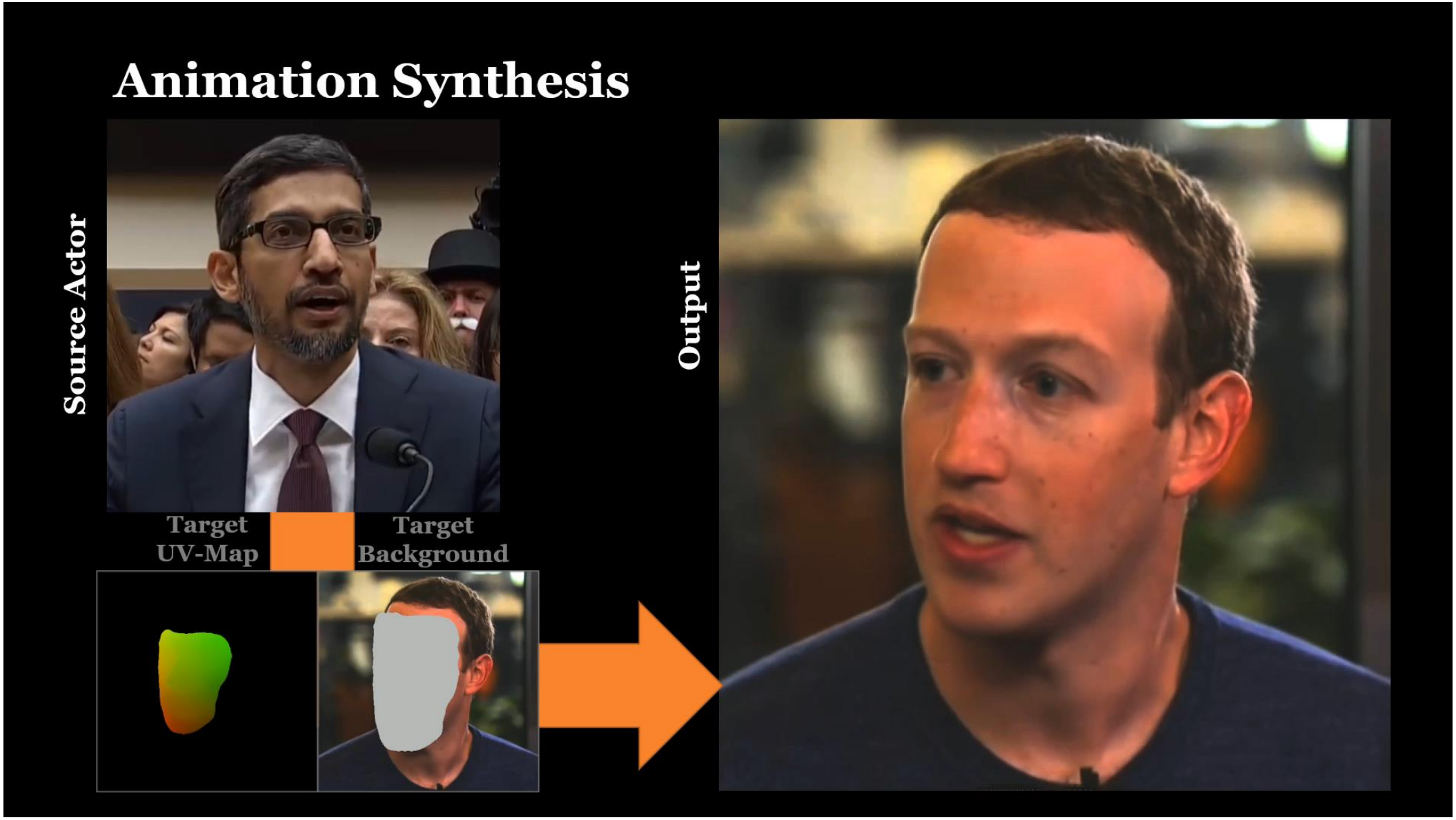
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



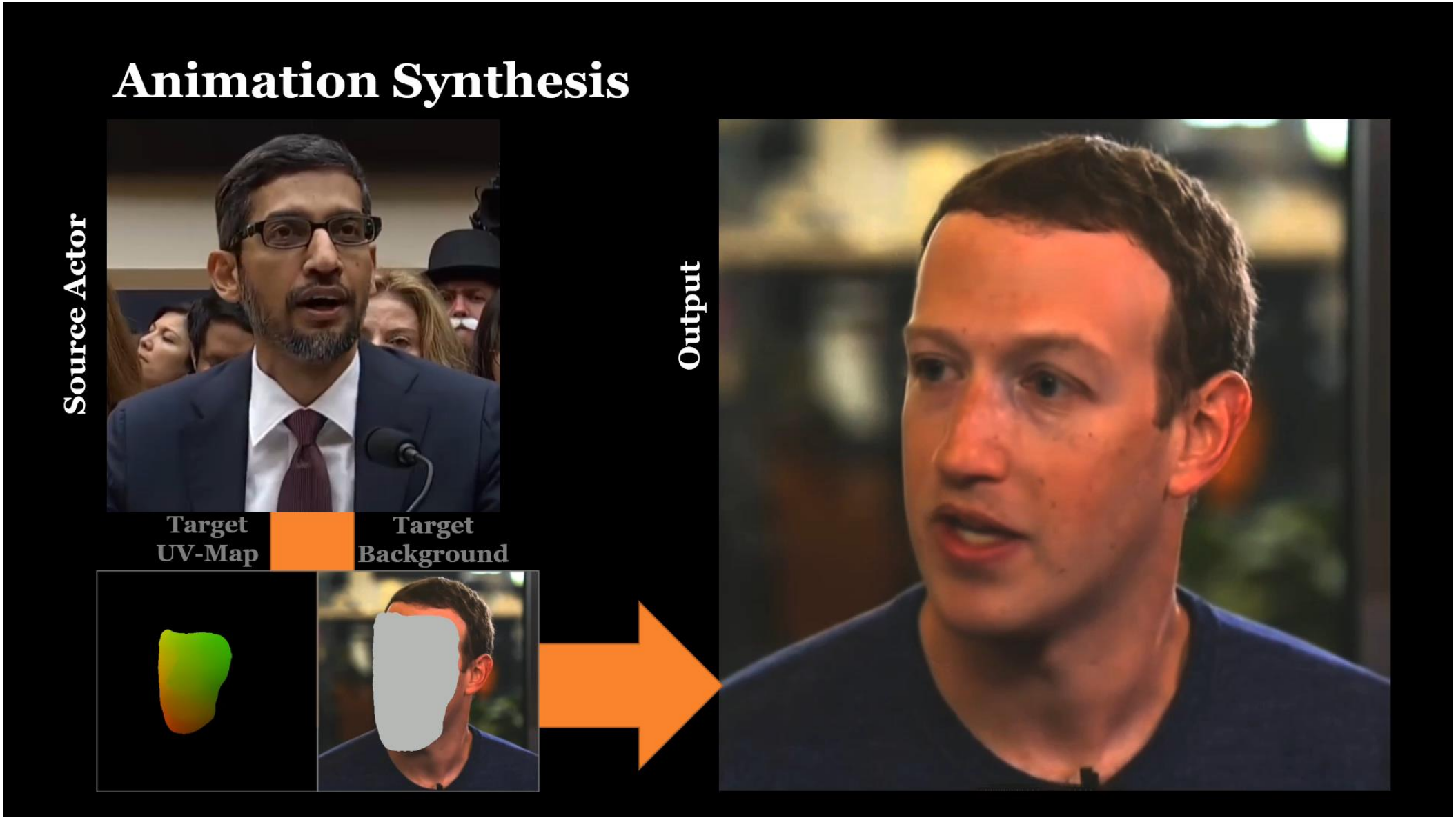
Siggraph'19 [Thies et al.]: Neural Textures

Deferred Neural Rendering



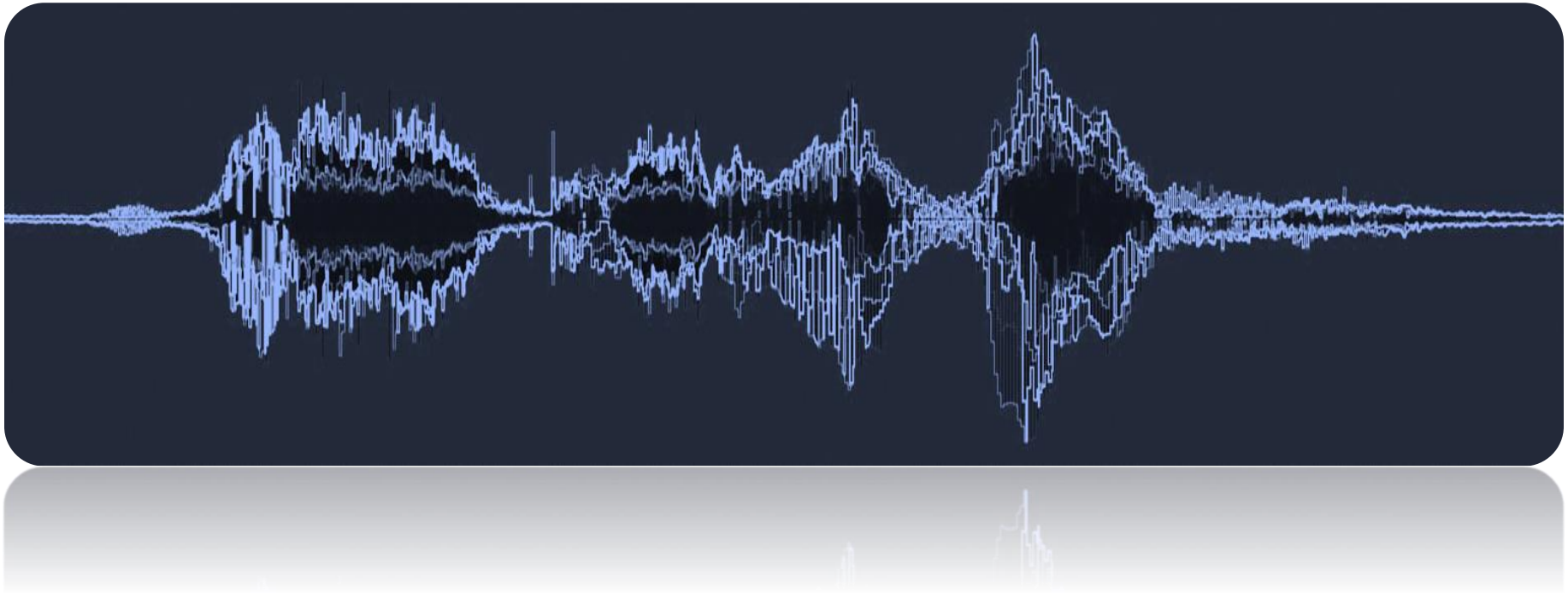
Siggraph'19 [Thies et al.]: Neural Textures

Deferred Neural Rendering



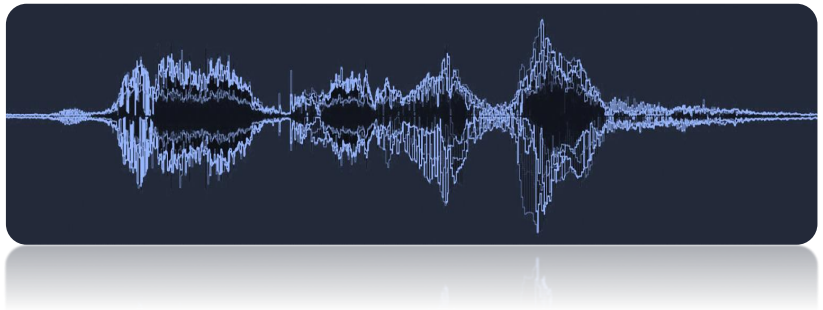
Siggraph'19 [Thies et al.]: Neural Textures

Conditioning on Audio: Neural Voice Puppetry



Audio to Video

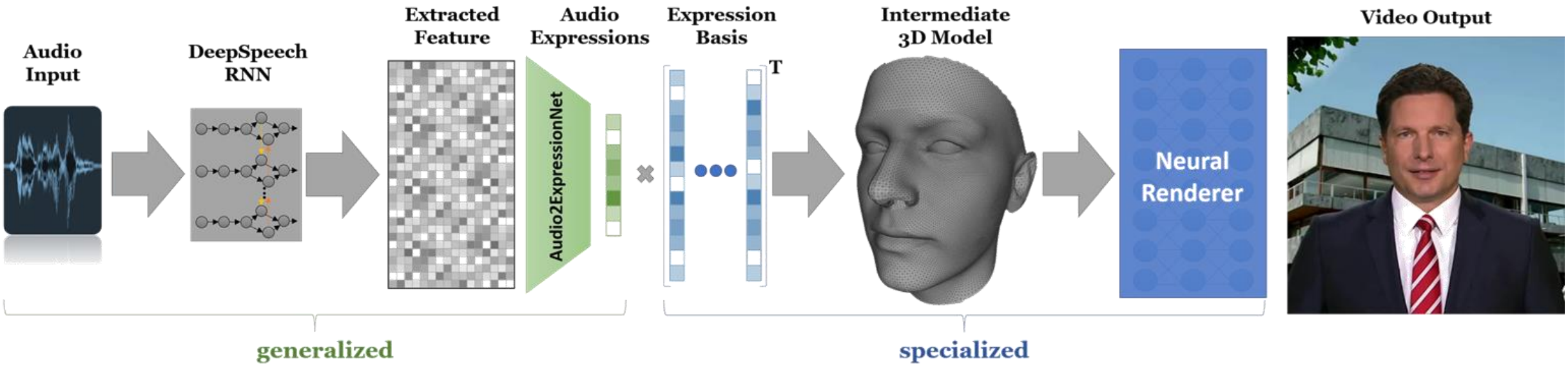
English Audio



German News Video

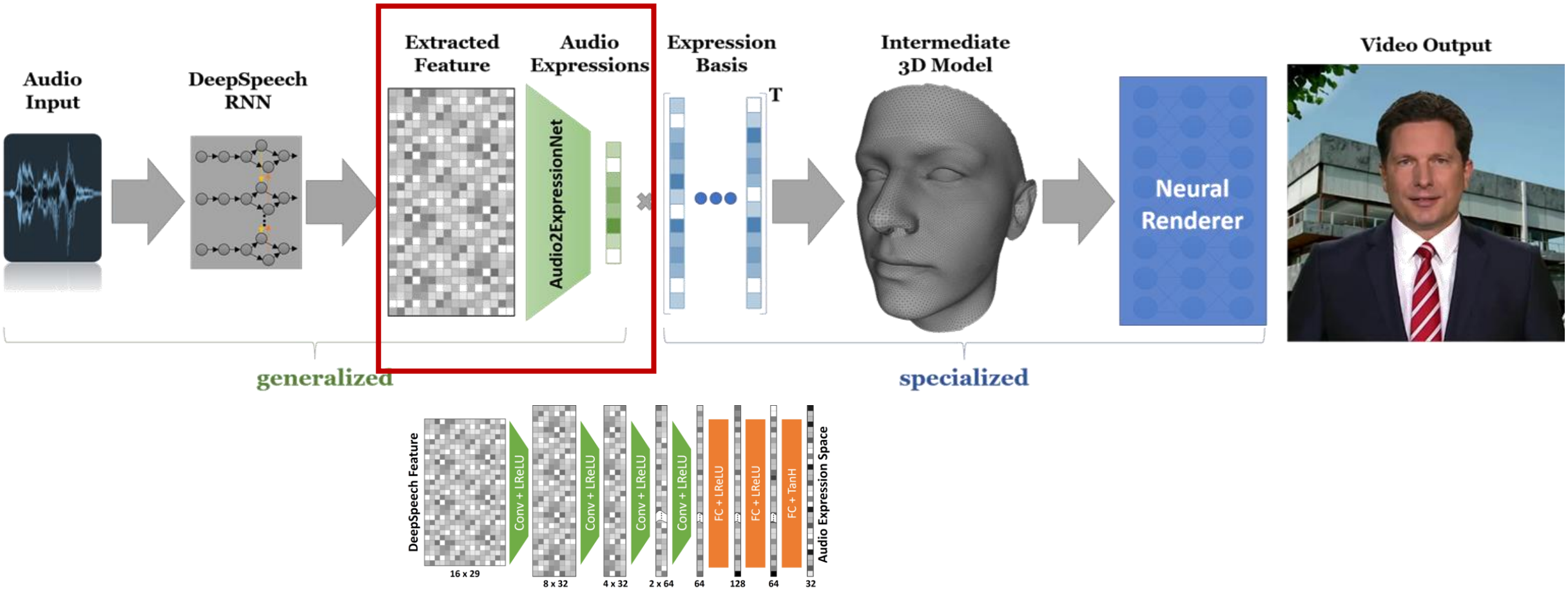
[Thies et al. 19]: Neural Voice Puppetry

Neural Voice Puppetry



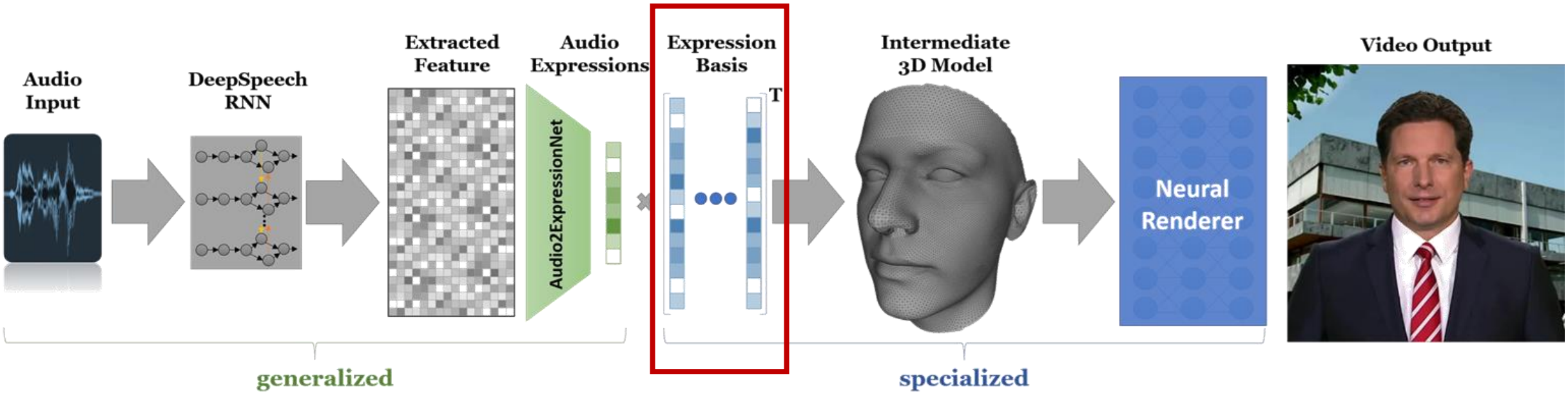
[Thies et al. 19]: Neural Voice Puppetry

Neural Voice Puppetry



[Thies et al. 19]: Neural Voice Puppetry

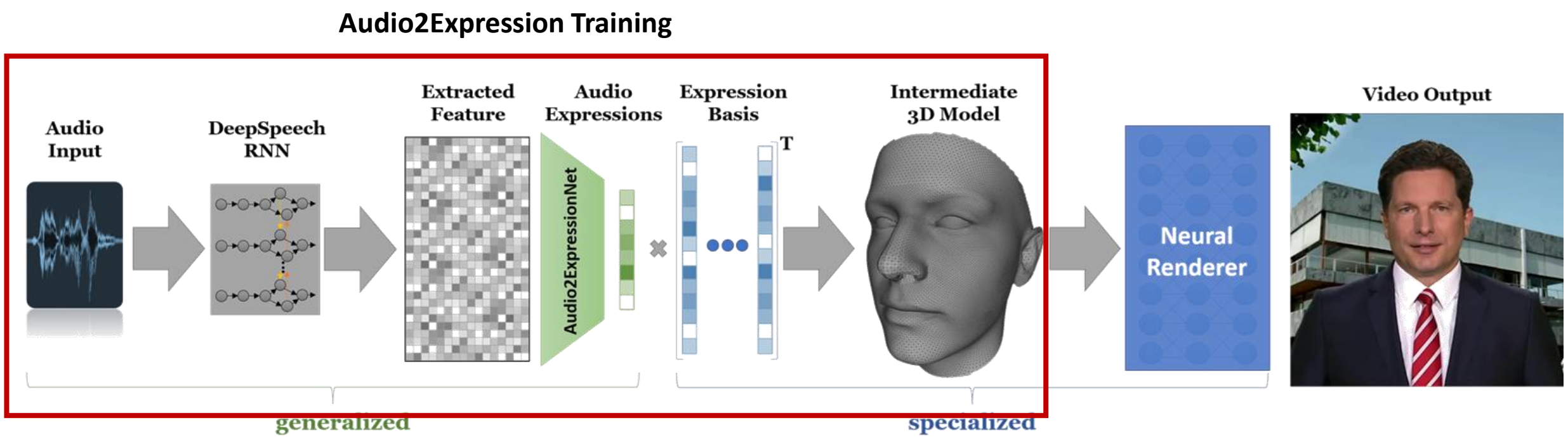
Neural Voice Puppetry



Person-specific
Blendshape Expression Model

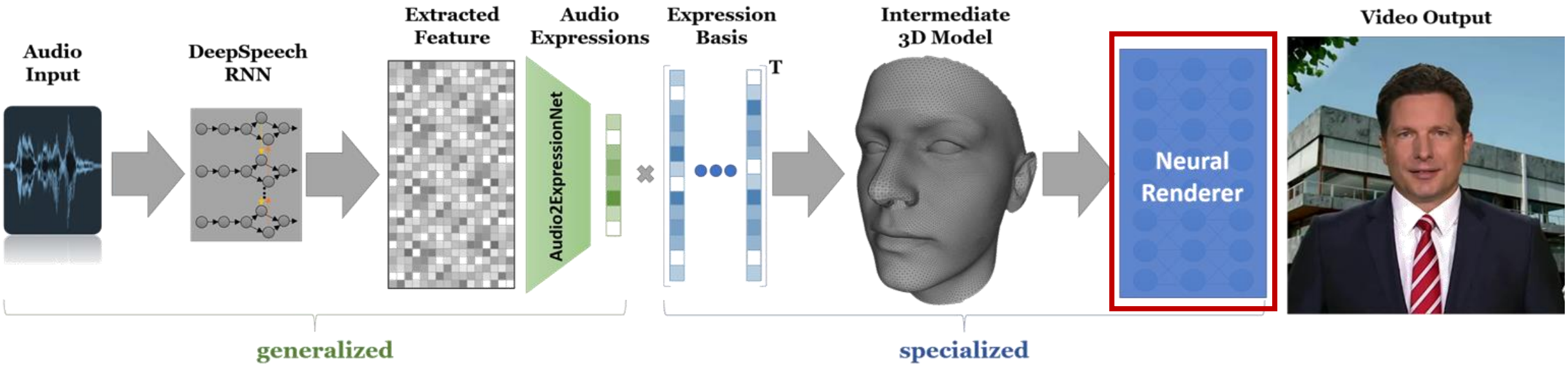
[Thies et al. 19]: Neural Voice Puppetry

Neural Voice Puppetry



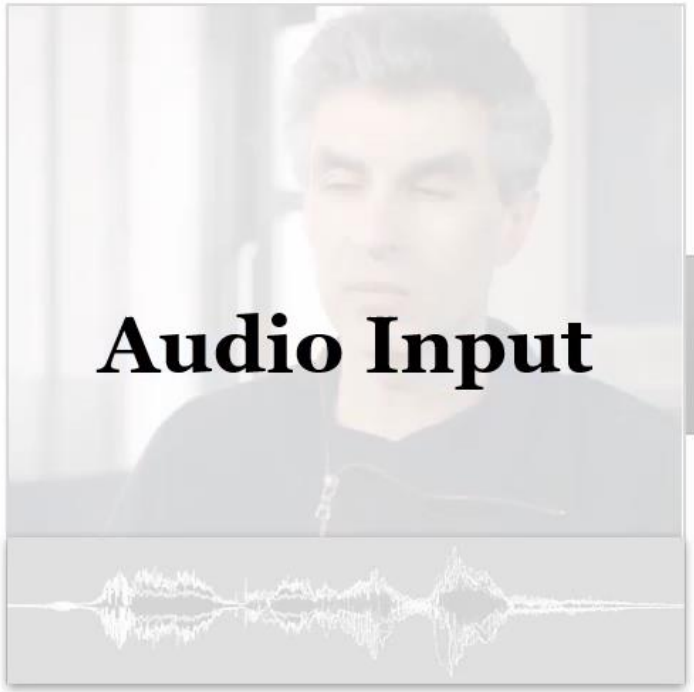
[Thies et al. 19]: Neural Voice Puppetry

Neural Voice Puppetry



[Thies et al. 19]: Neural Voice Puppetry

Neural Voice Puppetry: Audio to Video

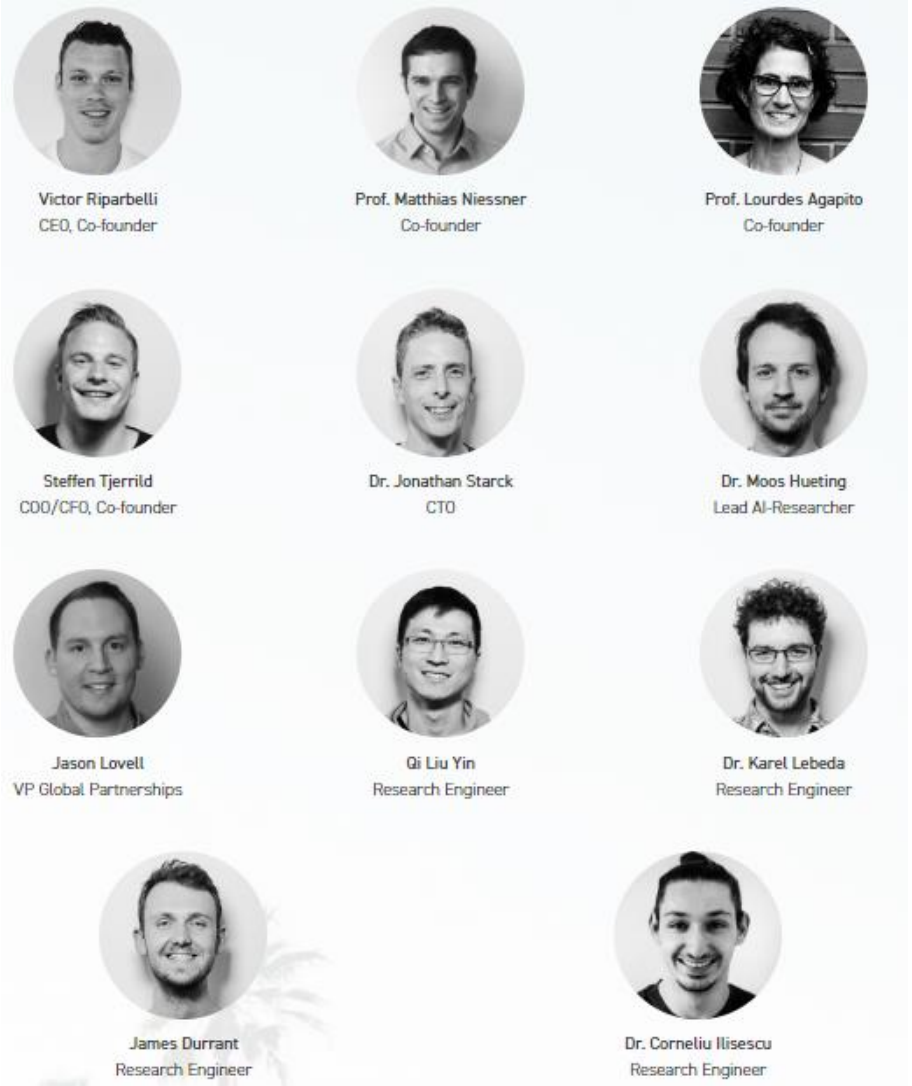


[Thies et al. 19]: Neural Voice Puppetry

Many Real-World Applications

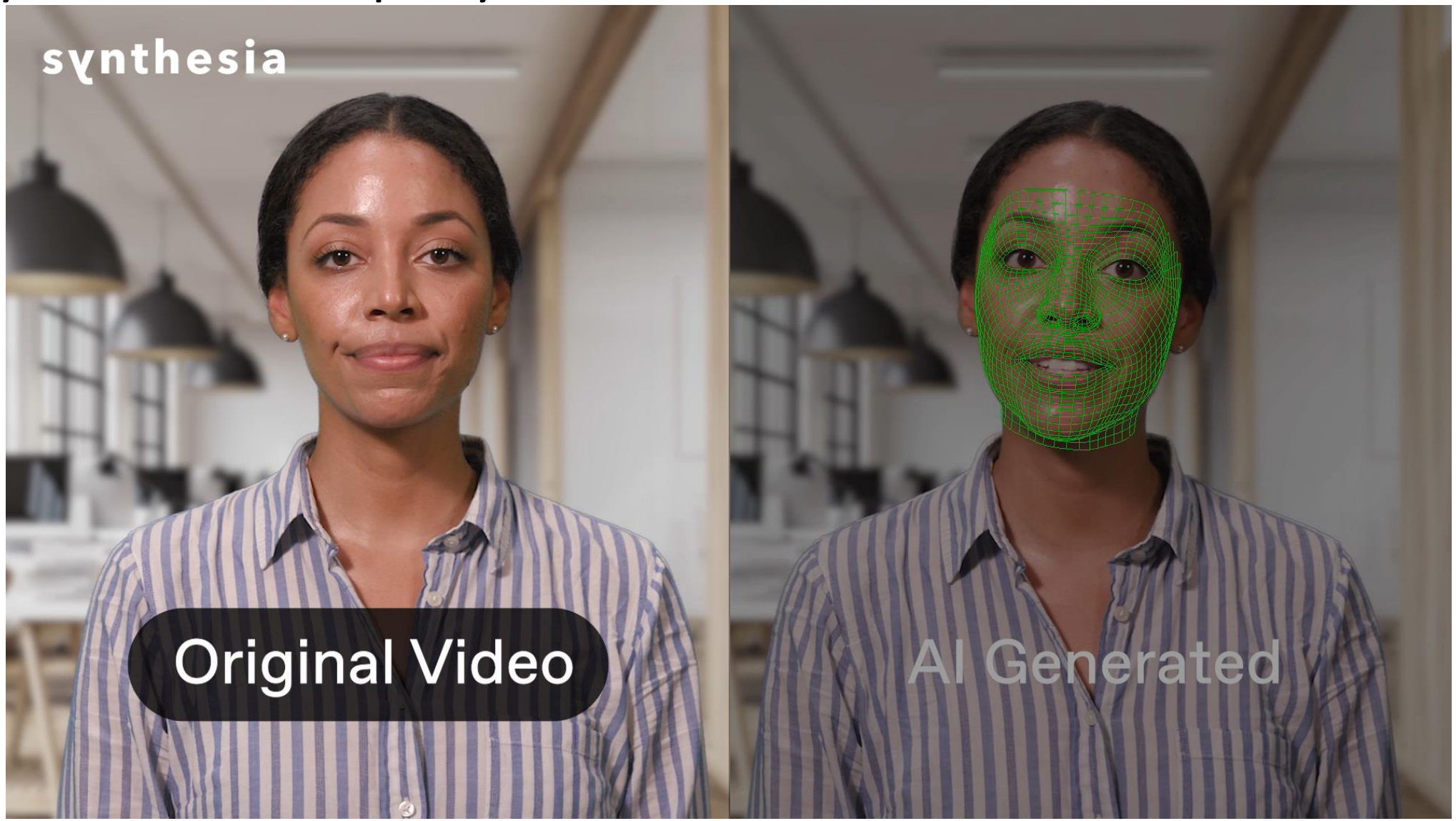
Synthesia: Lip Sync

synthesia



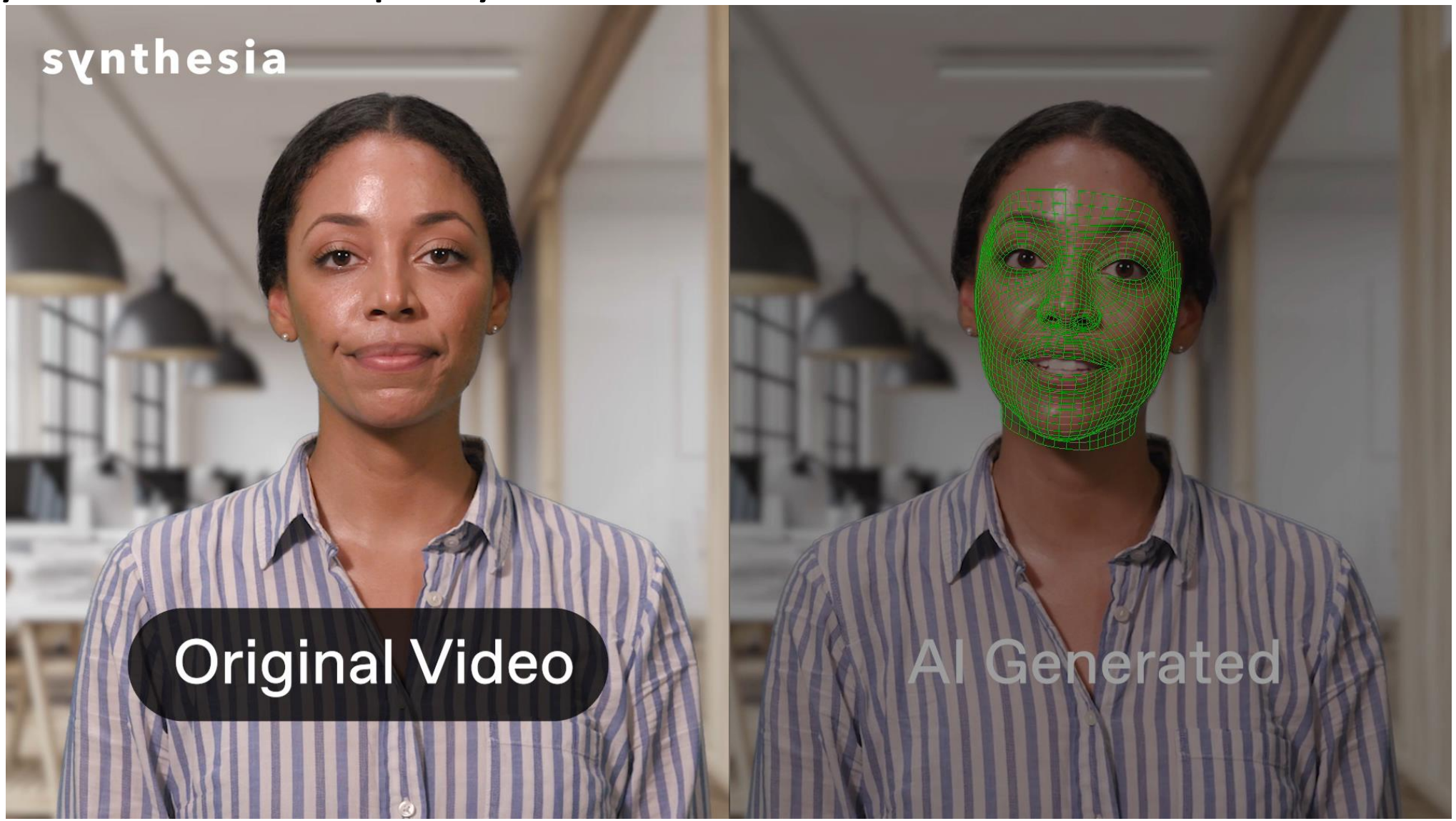
https://www.synthesia.io/

Synthesia: Lip Sync



https://www.synthesia.io/

Synthesia: Lip Sync



https://www.synthesia.io/

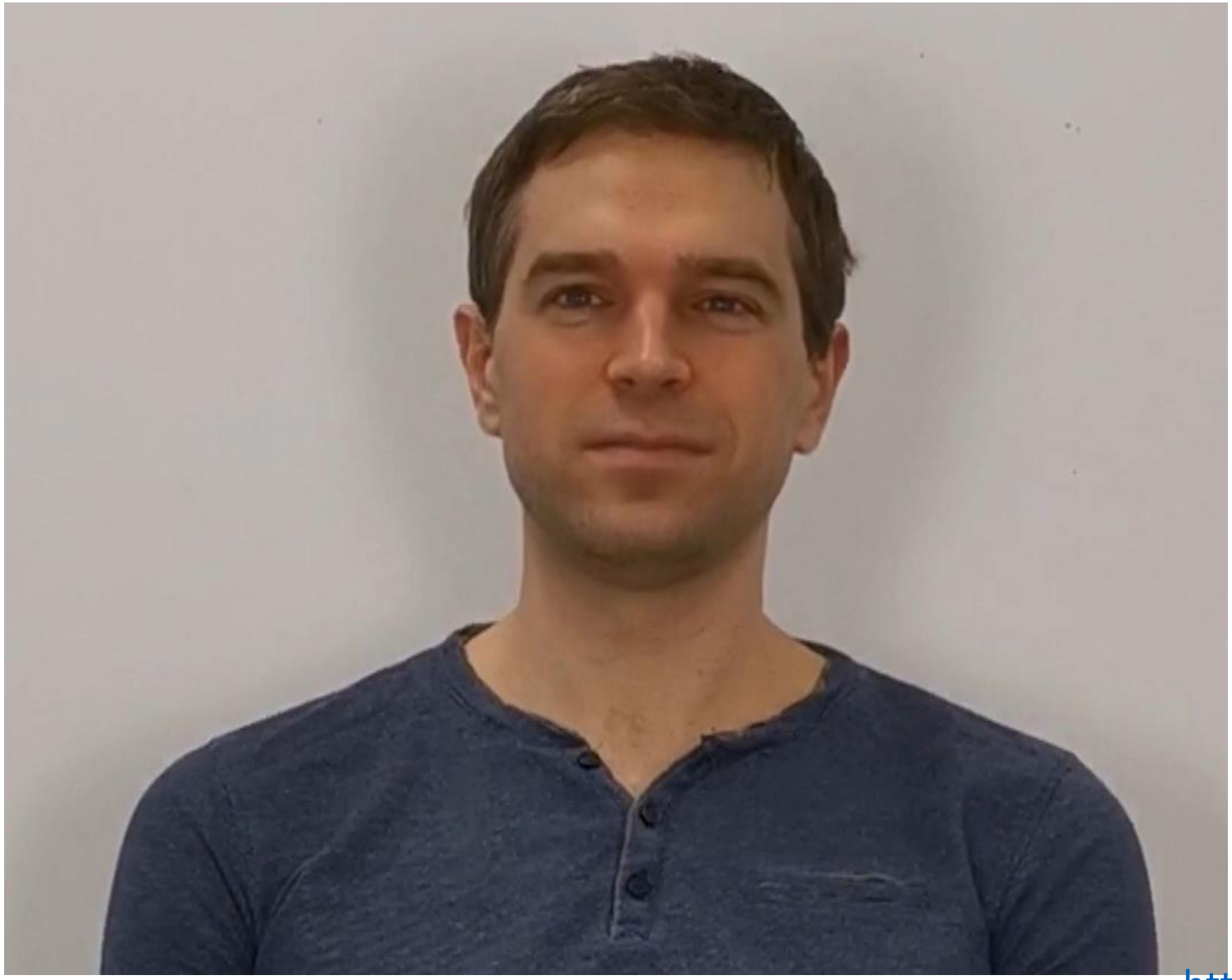
Synthesia Dubbing

https://www.malariamustdie.com/

Synthesia Dubbing

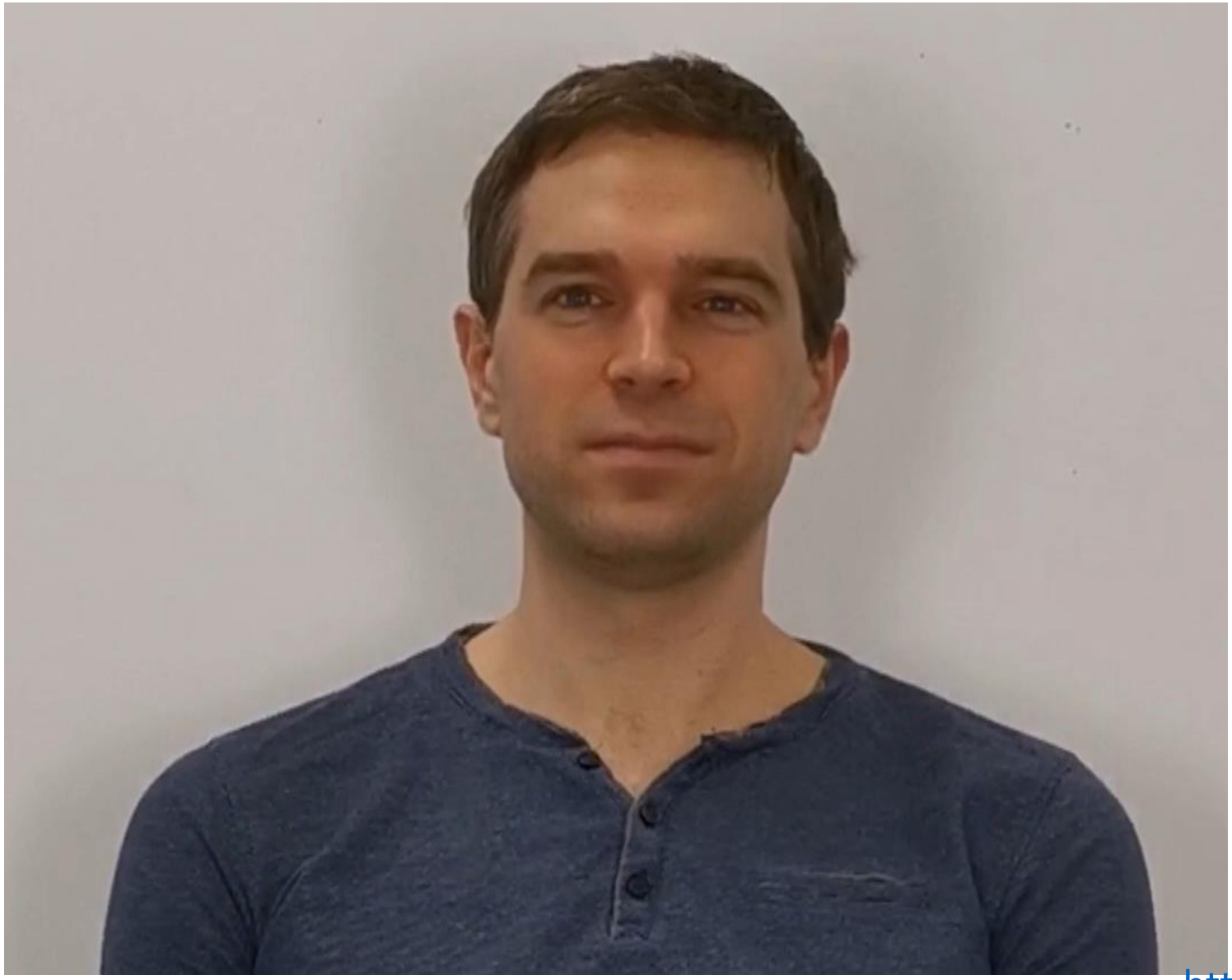
https://www.malariamustdie.com/

My Virtual Avatar



https://www.synthesia.io/

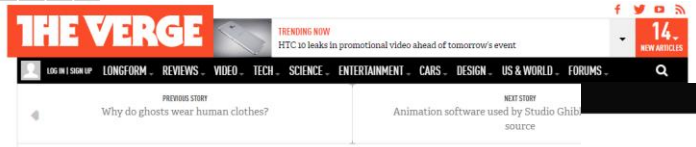
My Virtual Avatar

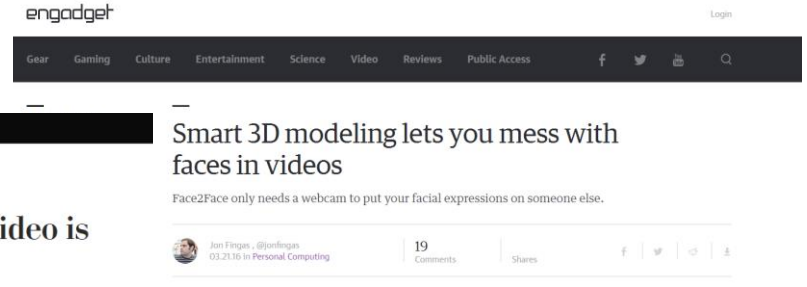


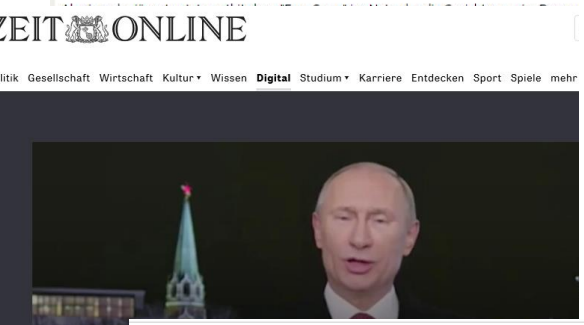
https://www.synthesia.io/

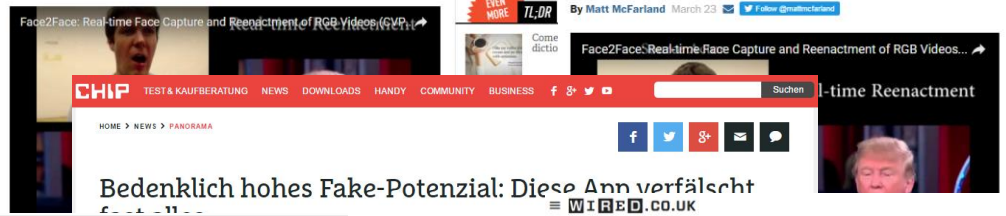
Video Editing is Popular

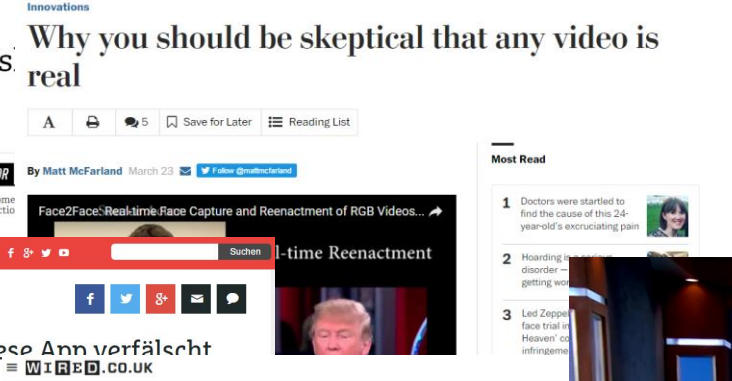
Video Editing is Popular

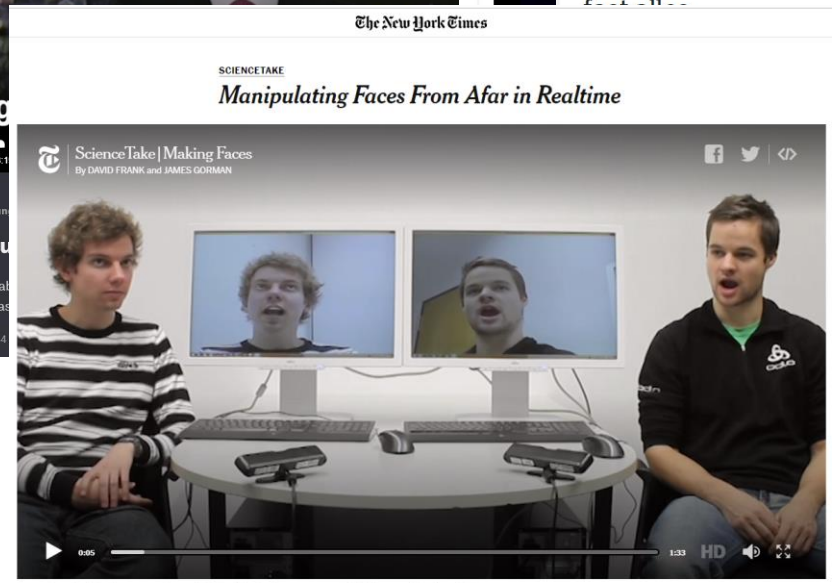












Computer scientists have a way to manipulate videos to change a person's facial expressions in real time.

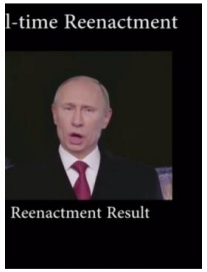


Bild: YouTube/Matthias Niessner

ssige Quelle zur Beweisführung
Verfälschungspotenzial, dass
stark an Wert verlieren könnten.

ANZI

Selected Press

Video Editing is very Popular

… and it’s fun!

IMU2Face

… and it’s fun!

IMU2Face

… and it’s fun!

Voice

NeuralVoicePuppetry

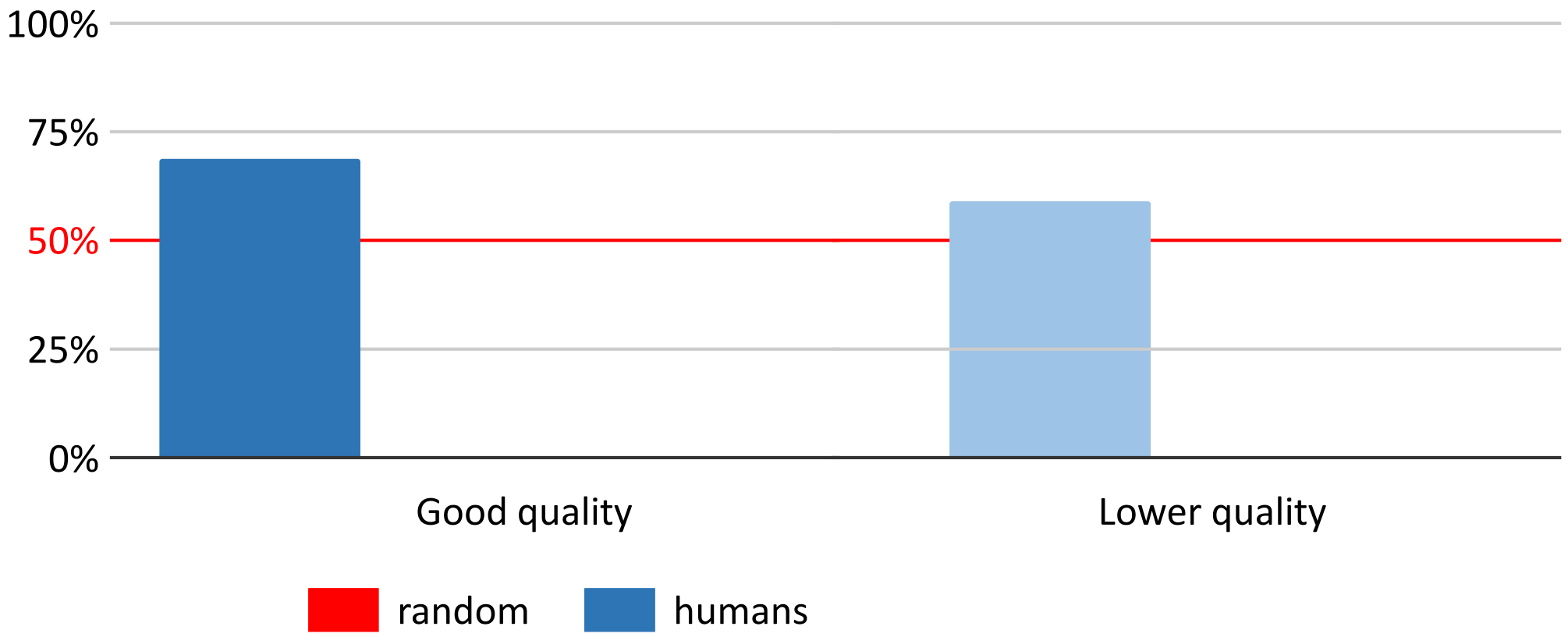
… and it’s fun!

Voice

NeuralVoicePuppetry

Need to think about ethics
and possible counter measures!

Study with over 200 participants



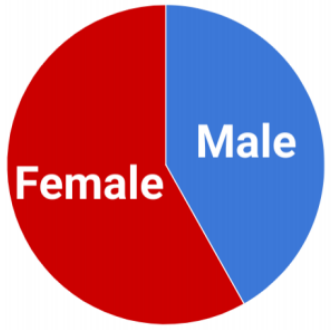
AI for Detection: FaceForensics

ICCV'19 [Roessler et al.]: FaceForensics++

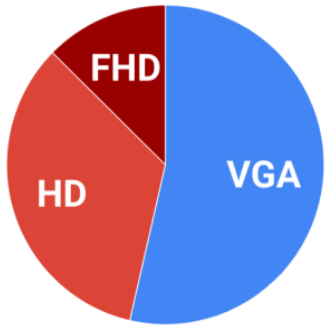
FaceForensics: Dataset

Source: 1,000 Videos (510,529 frames)

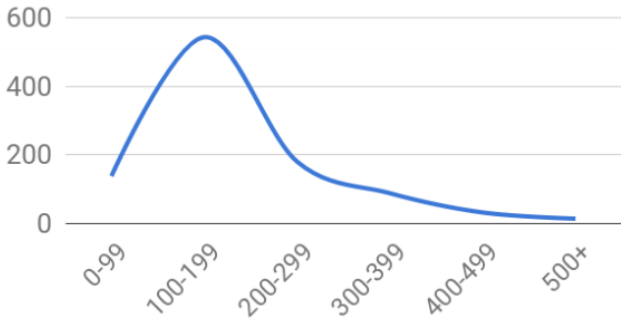
Methods	Train	Validation	Test
Pristine	366,847	68,511	73,770
DeepFakes	366,835	68,506	73,768
Face2Face	366,843	68,511	73,770
FaceSwap	291,434	54,618	59,640
NeuralTextures	291,834	54,630	59,672



(a) Gender



(b) Resolution



(c) Pixel Coverage of Faces

- Publicly available!
- Over 2 million manipulated frames
- Three compression levels for each manipulated frame
- Over 1500 research groups

ICCV'19 [Roessler et al.]: FaceForensics++

FaceForensics: Deep Detection Dataset

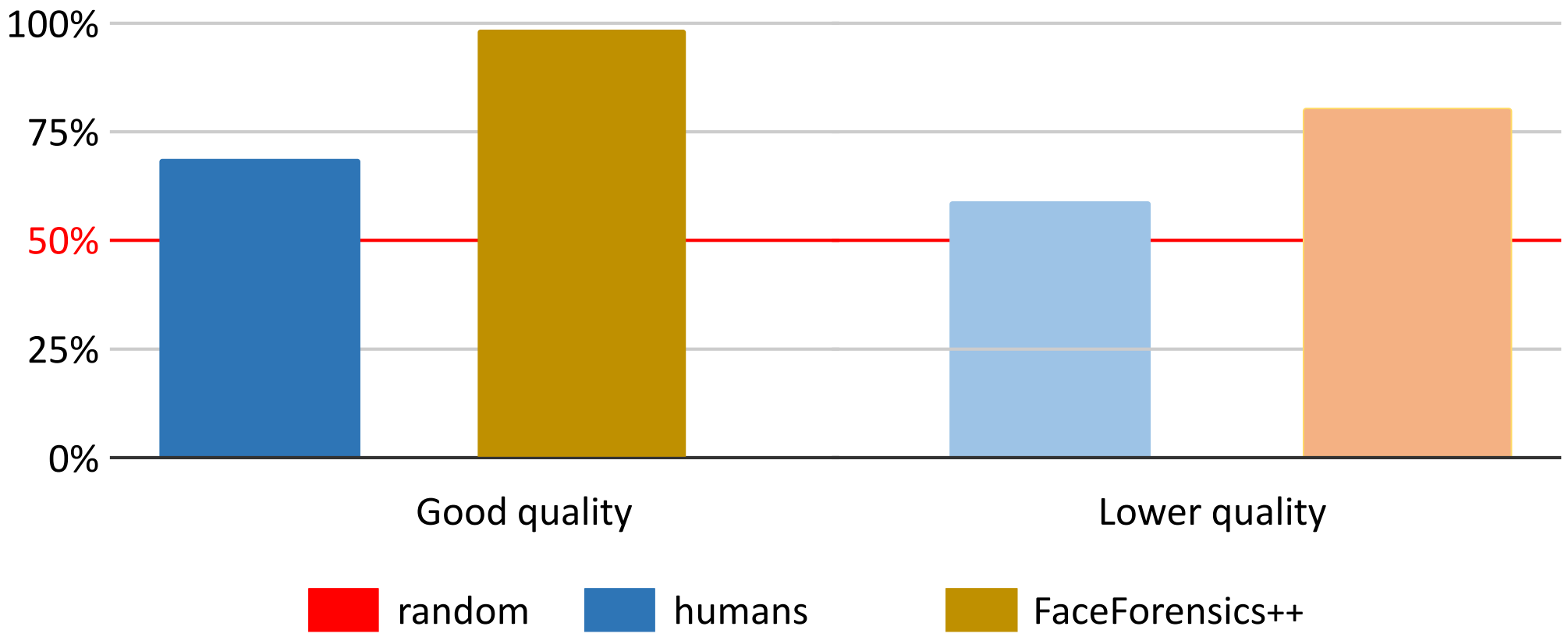
- Over 3000 manipulated videos
- From 28 actors
- Variety of scenes
- Provided by Google & JigSaw

ICCV'19 [Roessler et al.]: FaceForensics++

FaceForensics: Supervised Detection

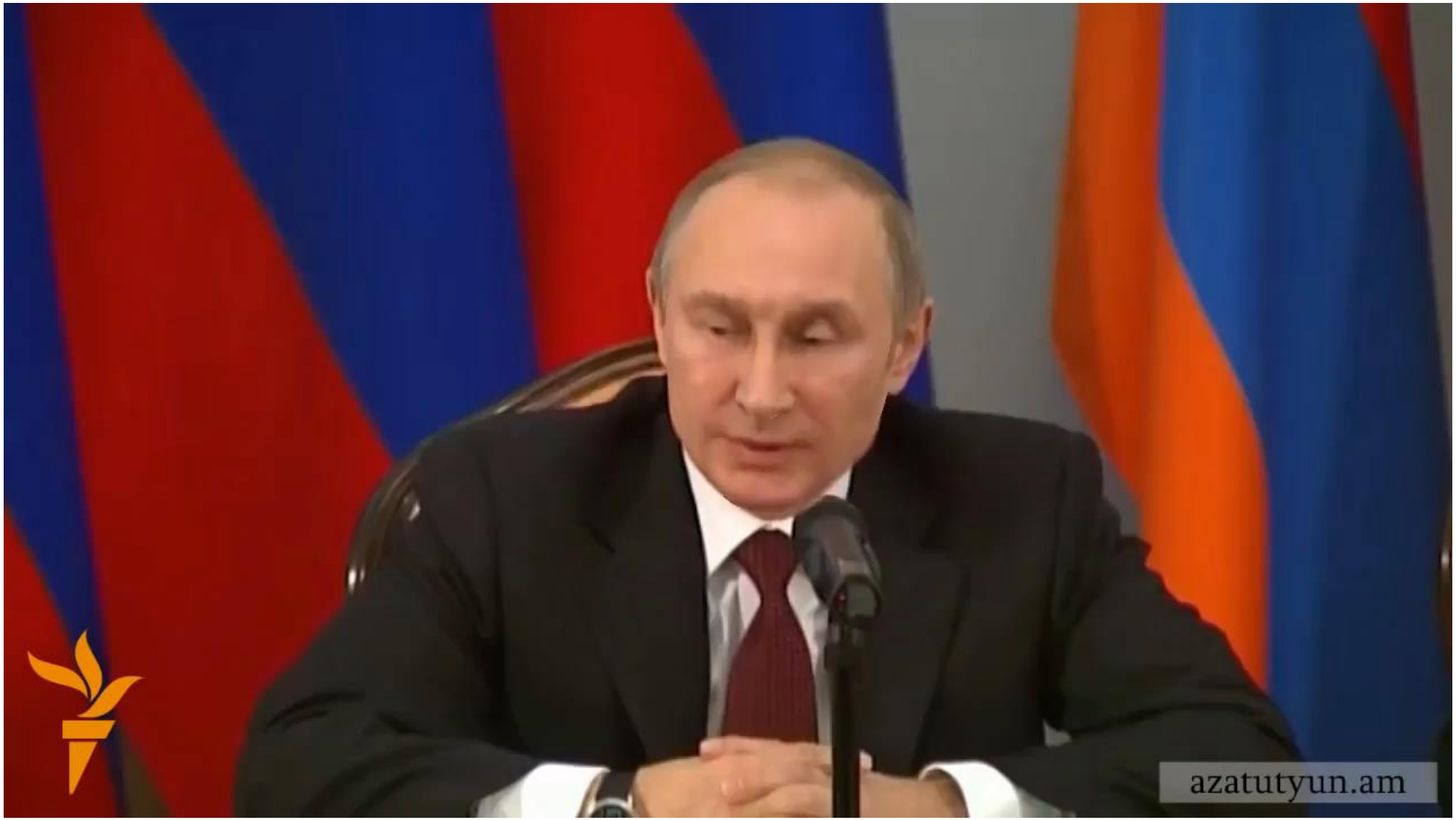
ICCV'19 [Roessler et al.]: FaceForensics++

FaceForensics++: Detection



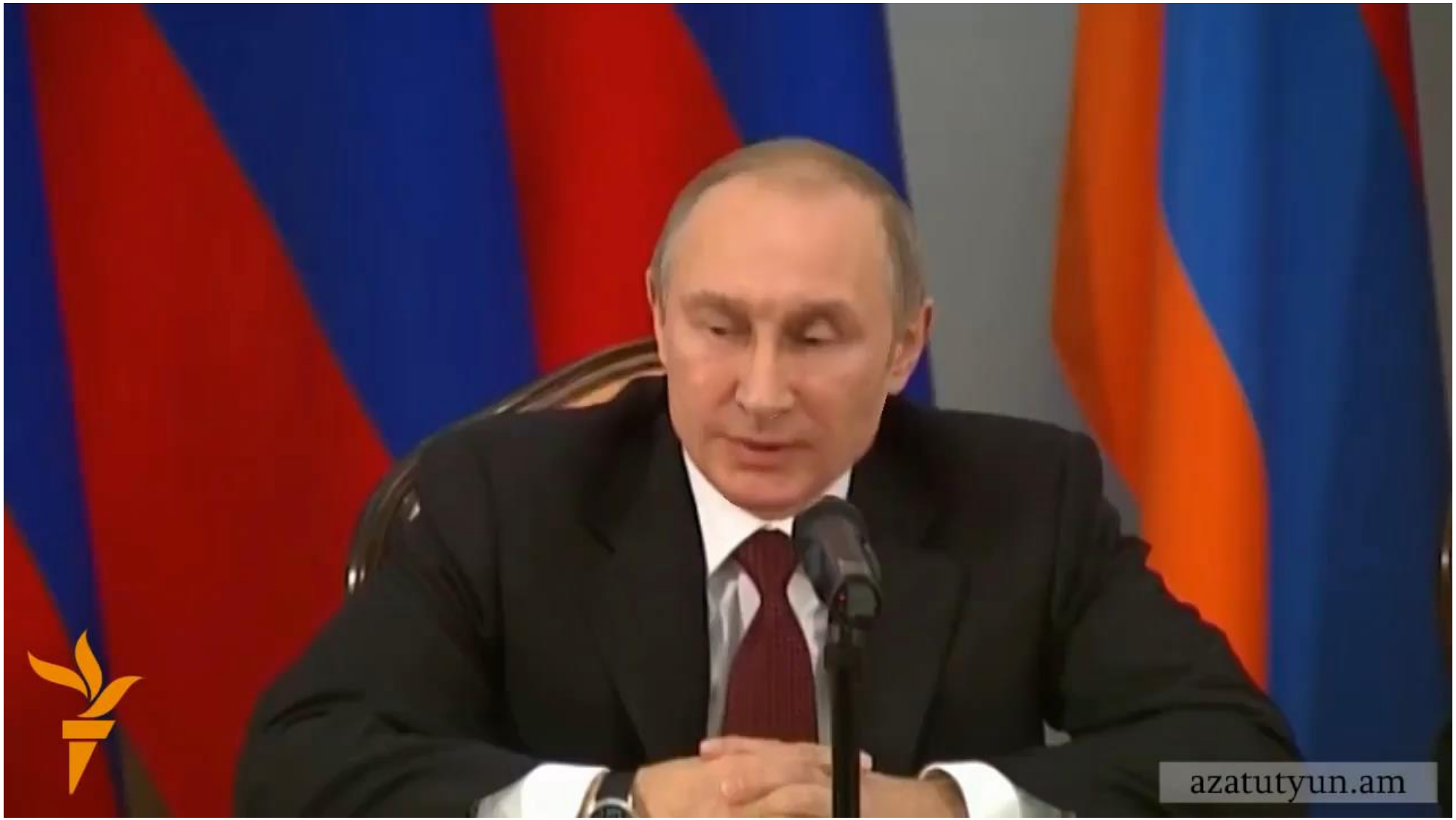
ICCV'19 [Roessler et al.]: FaceForensics++

FaceForensics++: Detection



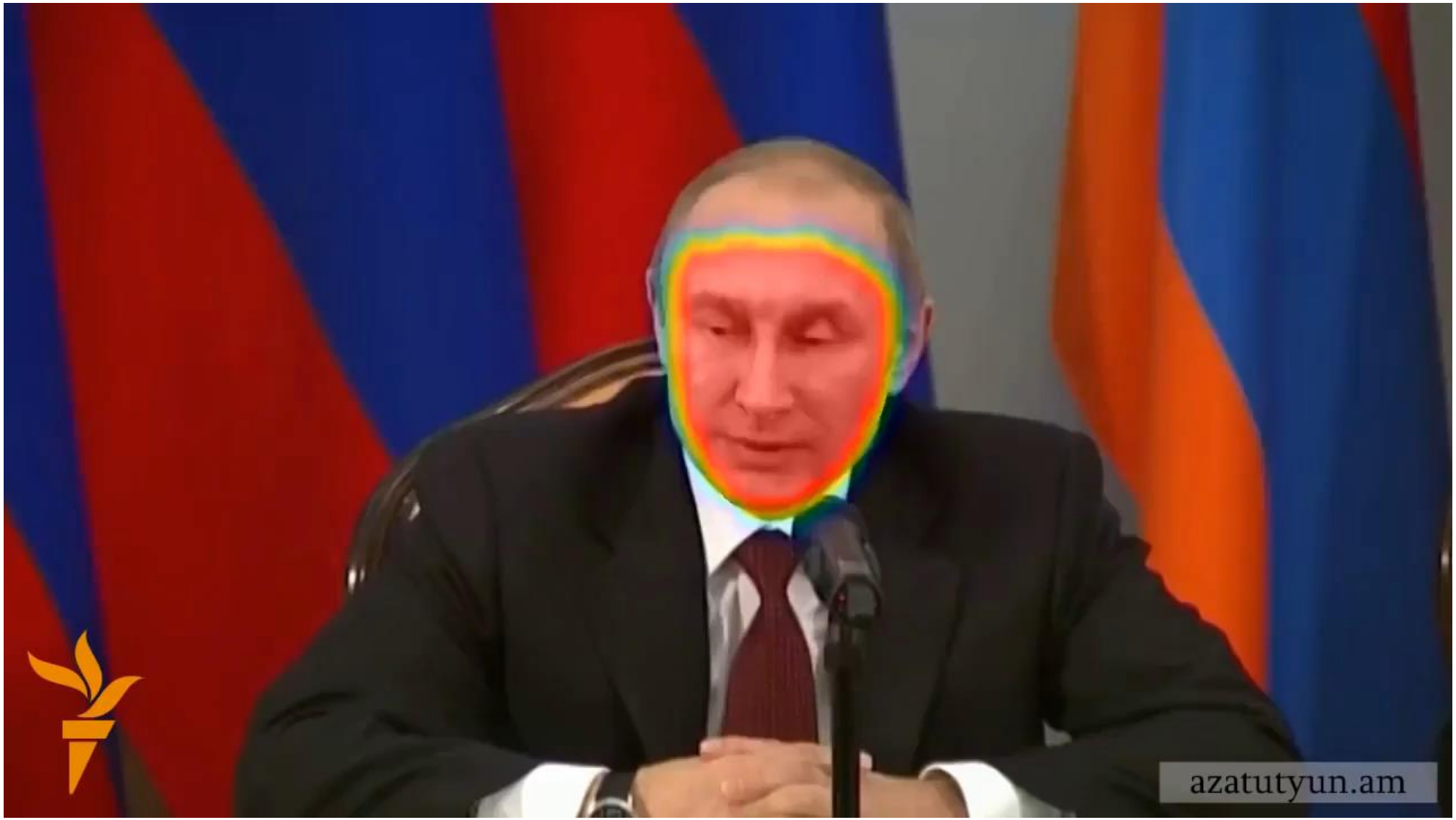
ICCV'19 [Roessler et al.]: FaceForensics++

FaceForensics++: Detection



ICCV'19 [Roessler et al.]: FaceForensics++

FaceForensics++: Detection

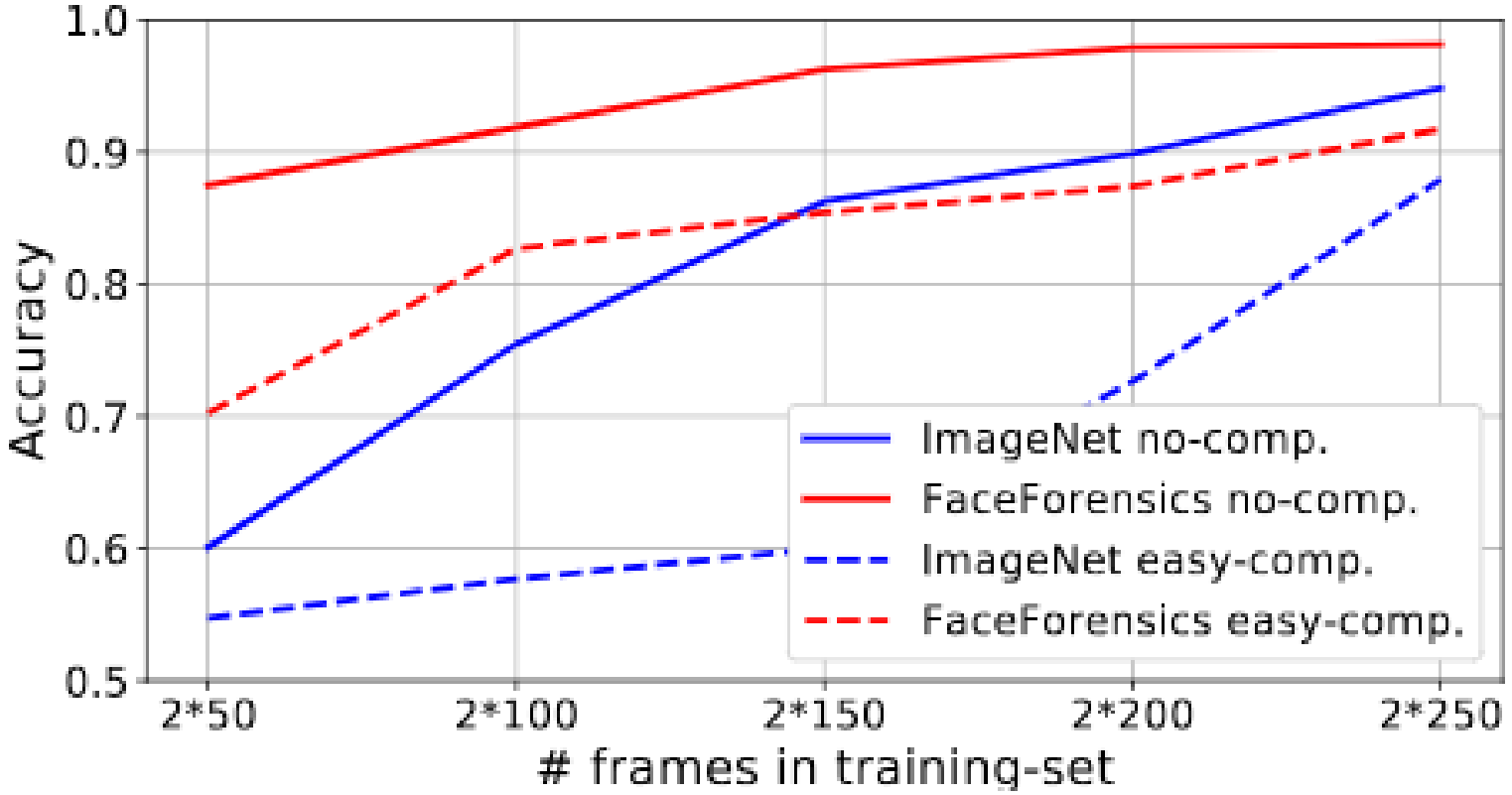


ICCV'19 [Roessler et al.]: FaceForensics++

Unsupervised / Self-Supervised Forensics

Major challenges

- Self-supervised Learning
- Transfer Learning
- Unsupervised Learning



[Cozzolino et al. 19]: ForensicTransfer

Conclusion

Synthesis

Synergy

Forensics

Vision == Inverse Graphics

Professor

Prof. Dr. Matthias Nießner

PostDocs

Dr. Justus Thies

Angel X. Chang
Visiting Professor:
Hans Fischer Fellow
Princeton University

Leonidas Guibas
Visiting Professor:
Hans Fischer Senior Fellow
Stanford University

PhD Students

Aljaž Božič

Andreas Rössler

Armen Avetisyan

Ji Hou

Dejan Azinović

Manuel Dahnert

Dave Zhenyu Chen

Norman Müller

Yawar Siddiqui

TUM VISUAL COMPUTING

Thank You!

Visual Computing Group
Prof. Matthias Nießner

Thank You!

Visual Computing Group
Prof. Matthias Nießner