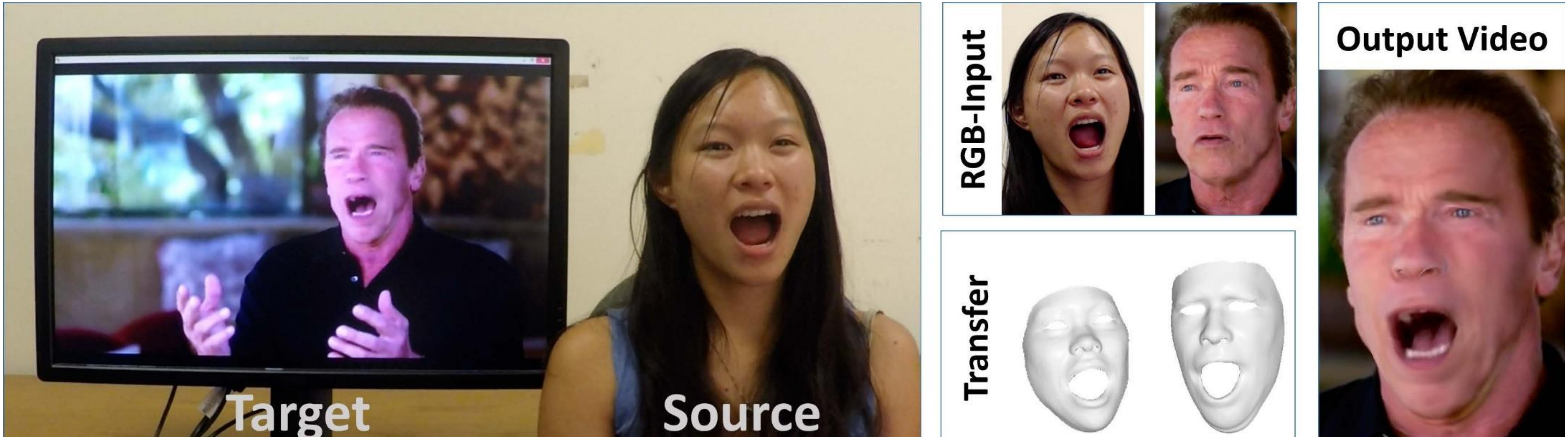
DeepFake Generation and Detection



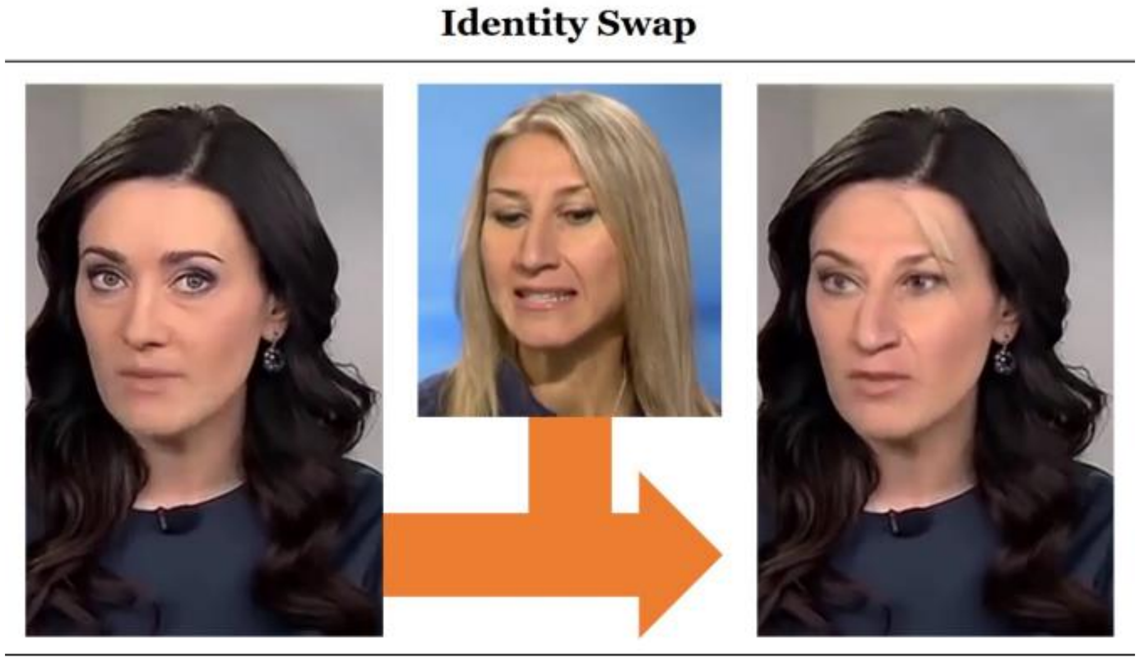
Visual Computing &
Artificial Intelligence

Prof. Matthias Nießner

The ‘Original’ DeepFake Method

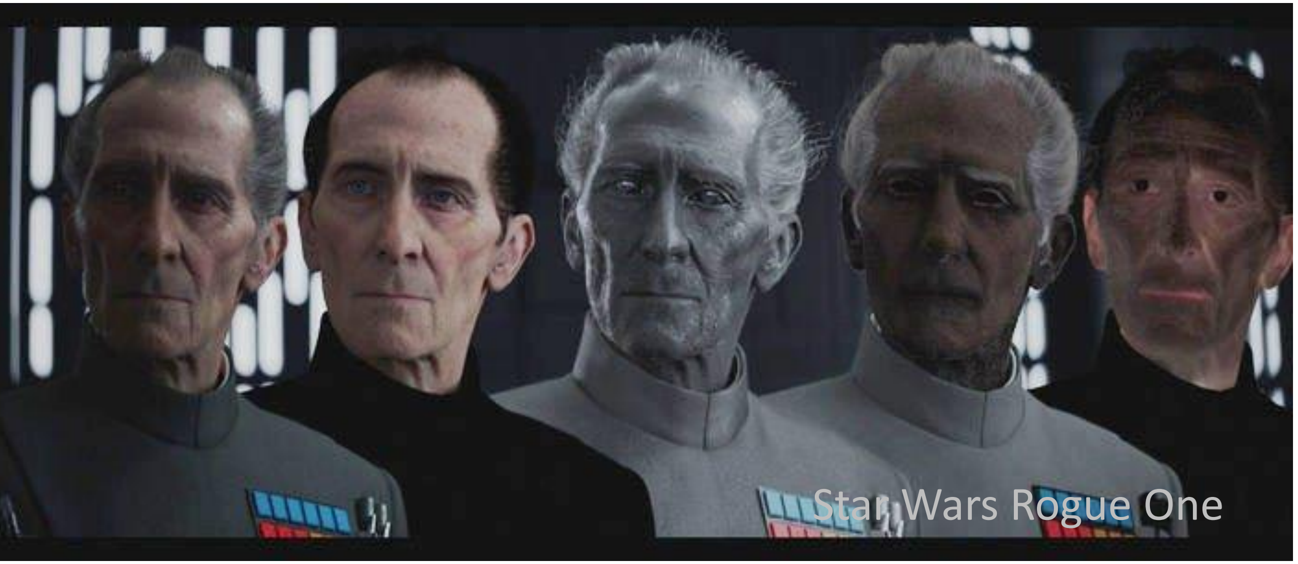
https://github.com/deepfakes/faceswap (github account name)

Face Swap vs Reenactment



Graphics vs Deep Learning

3D Model + Textures + Shading -> Synthetic Image



Generative Adversarial Networks

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

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

Graphics-based Facial Editing

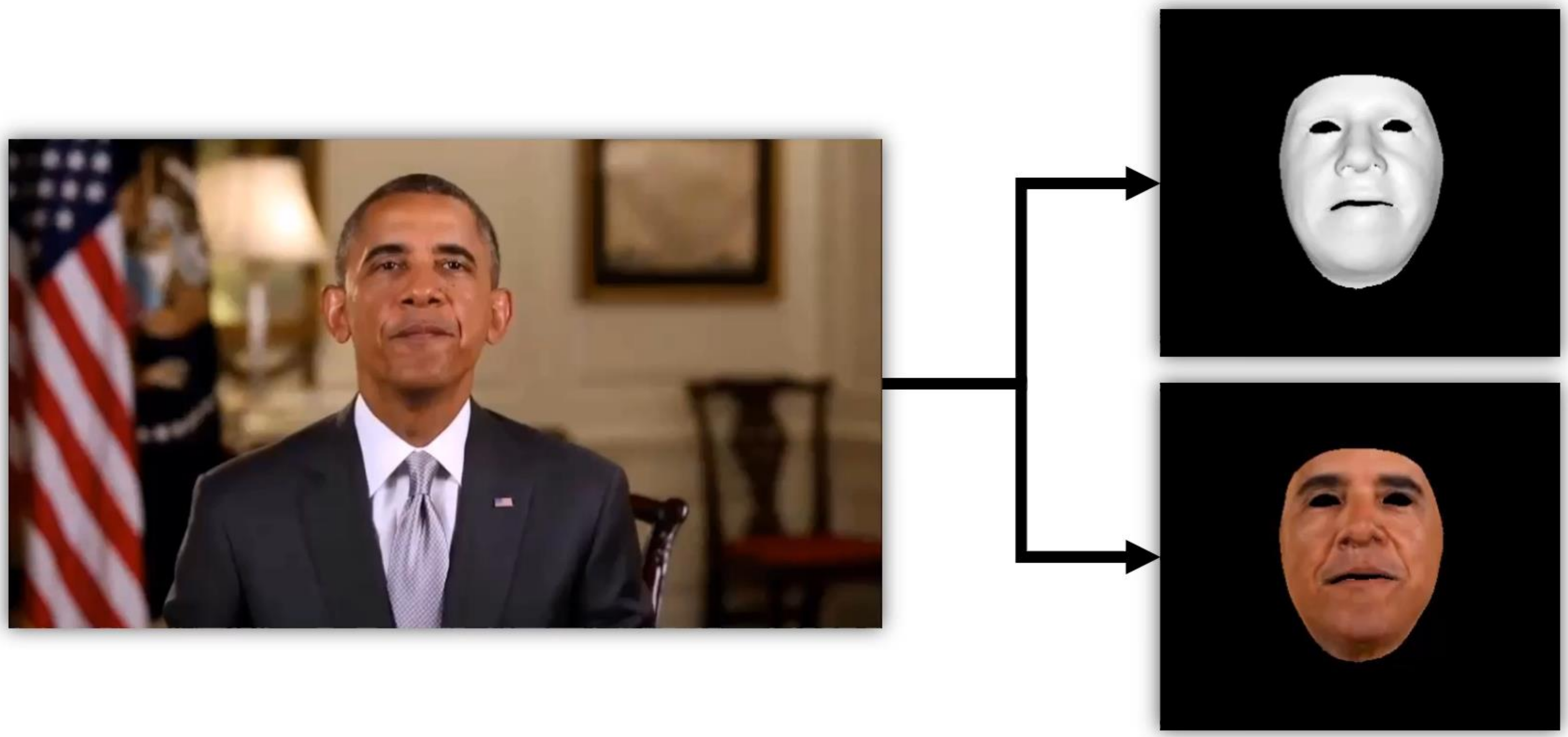
Fitting Parametric Model to RGB Image

$E(P) =$

$-$

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

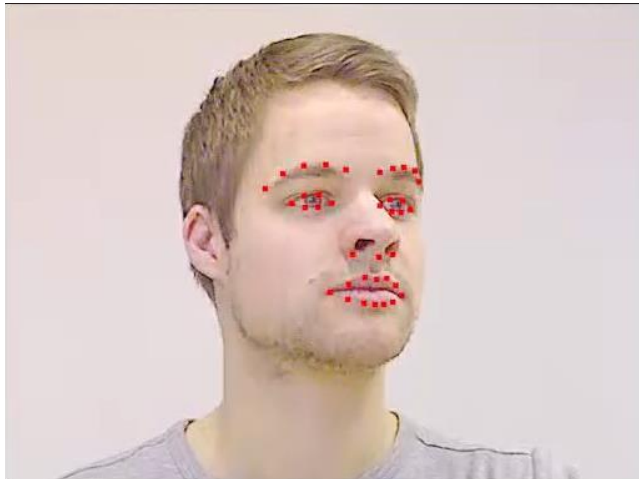
Fitting Parametric Model to RGB Image



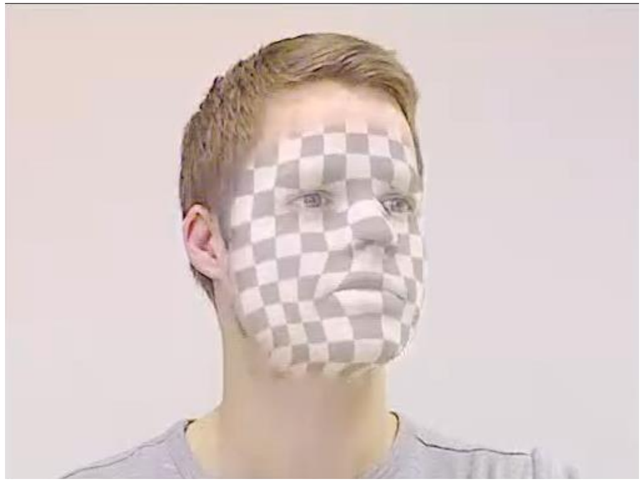
CVPR’16 (Oral) [Thies et al.]: Face2Face

3D Model + Image-based Rendering

Input

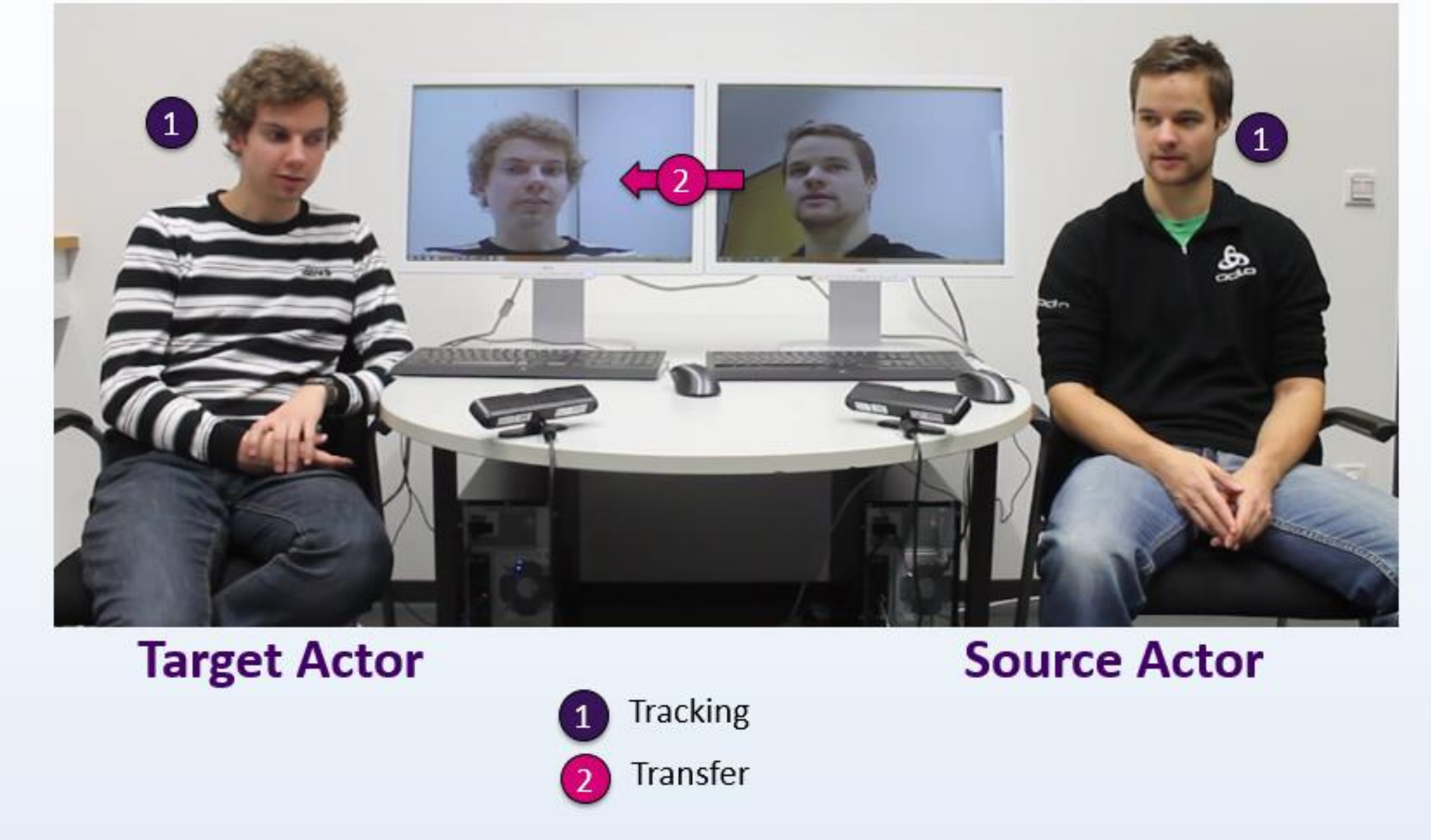


Output



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

Siggraph'18 [Thies et al.]: HeadOn

HeadOn: Reenactment of Portrait Videos

Siggraph'18 [Thies et al.]: HeadOn

DeepLearning-based Facial Editing

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*

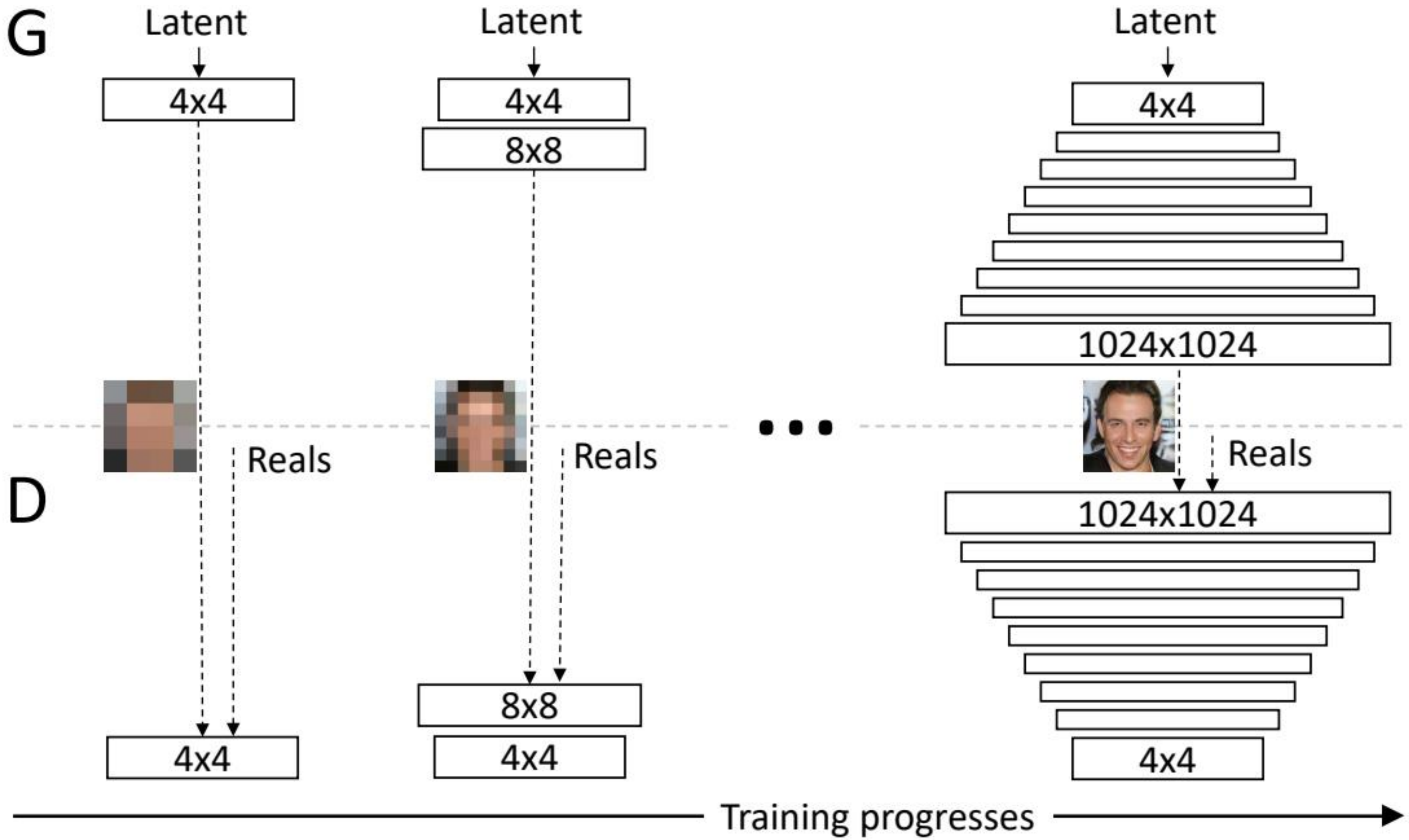
No explicit no control
-> *struggle with videos*

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

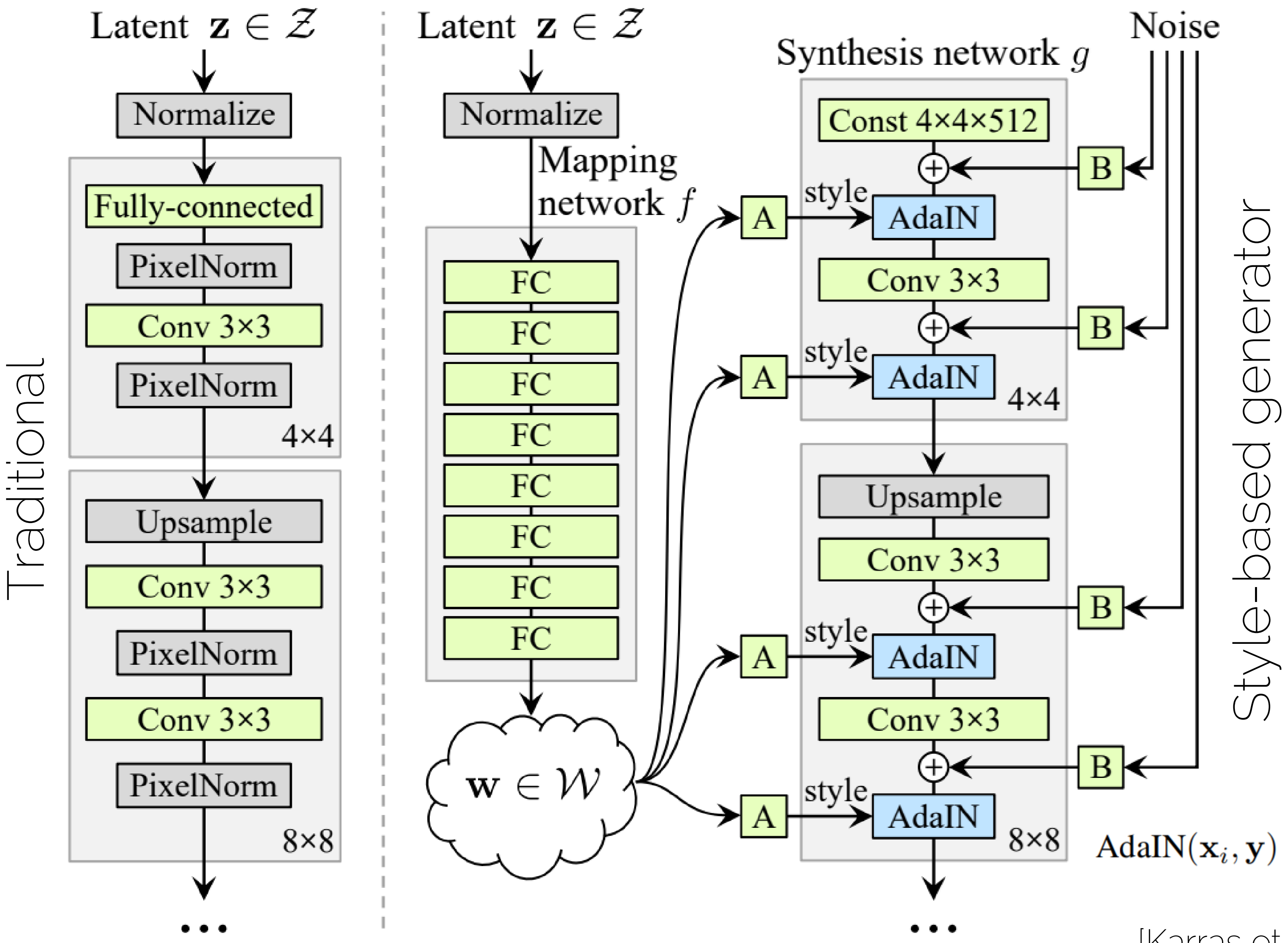
GANs [Goodfellow et al. 14], Pix2Pix [Isola et al. 17], ProGAN [Karras et al. 18], …

Progressive Growing GANs



https://github.com/tkarras/progressive_growing_of_gans [Karras et al. 17]

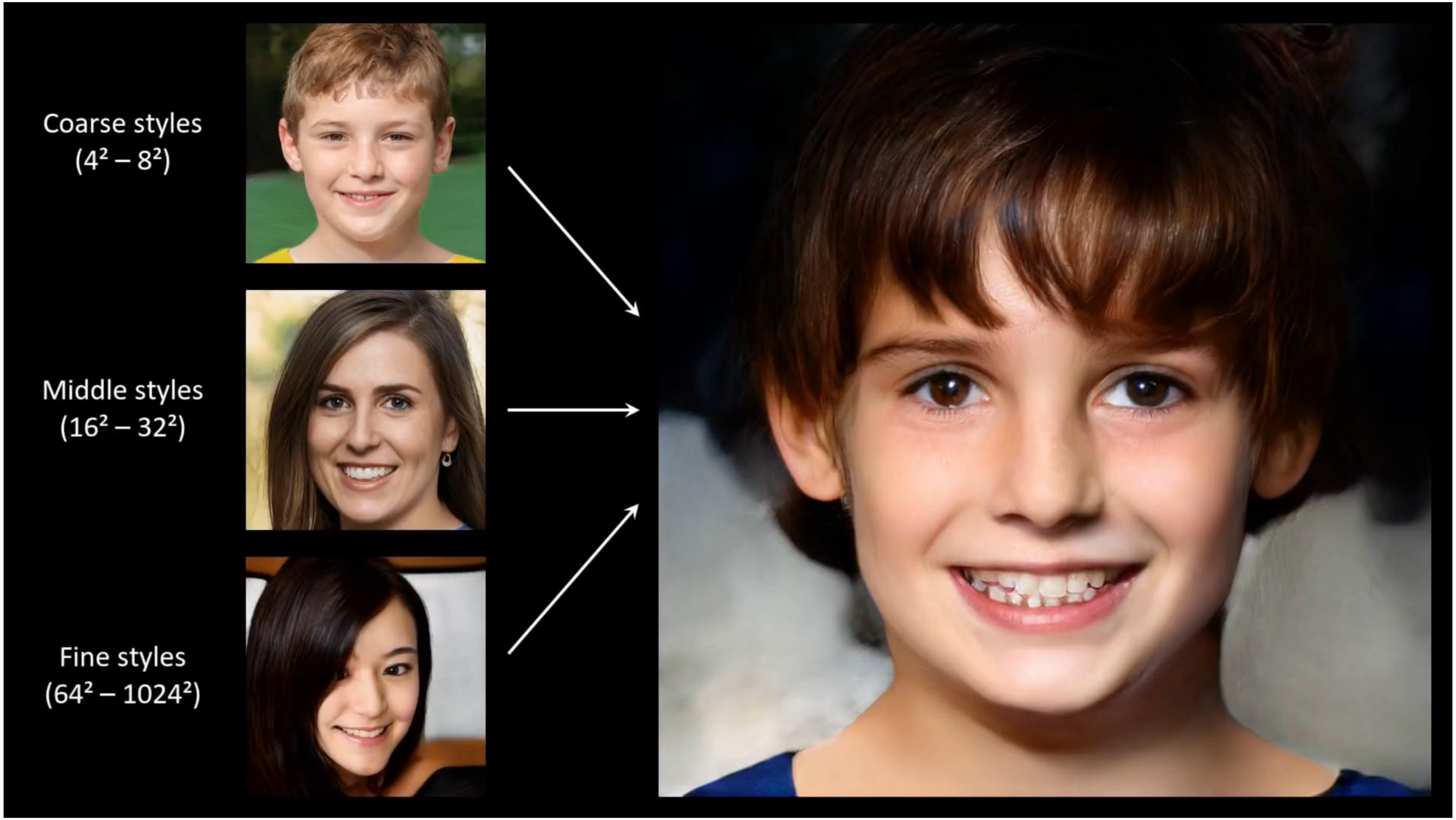
StyleGAN



$$\mathrm{AdaIN}(\mathbf{x}_i, \mathbf{y}) = \mathbf{y}_{s,i} \frac{\mathbf{x}_i - \mu(\mathbf{x}_i)}{\sigma(\mathbf{x}_i)} + \mathbf{y}_{b,i},$$

[Karras et al. 19] StyleGAN

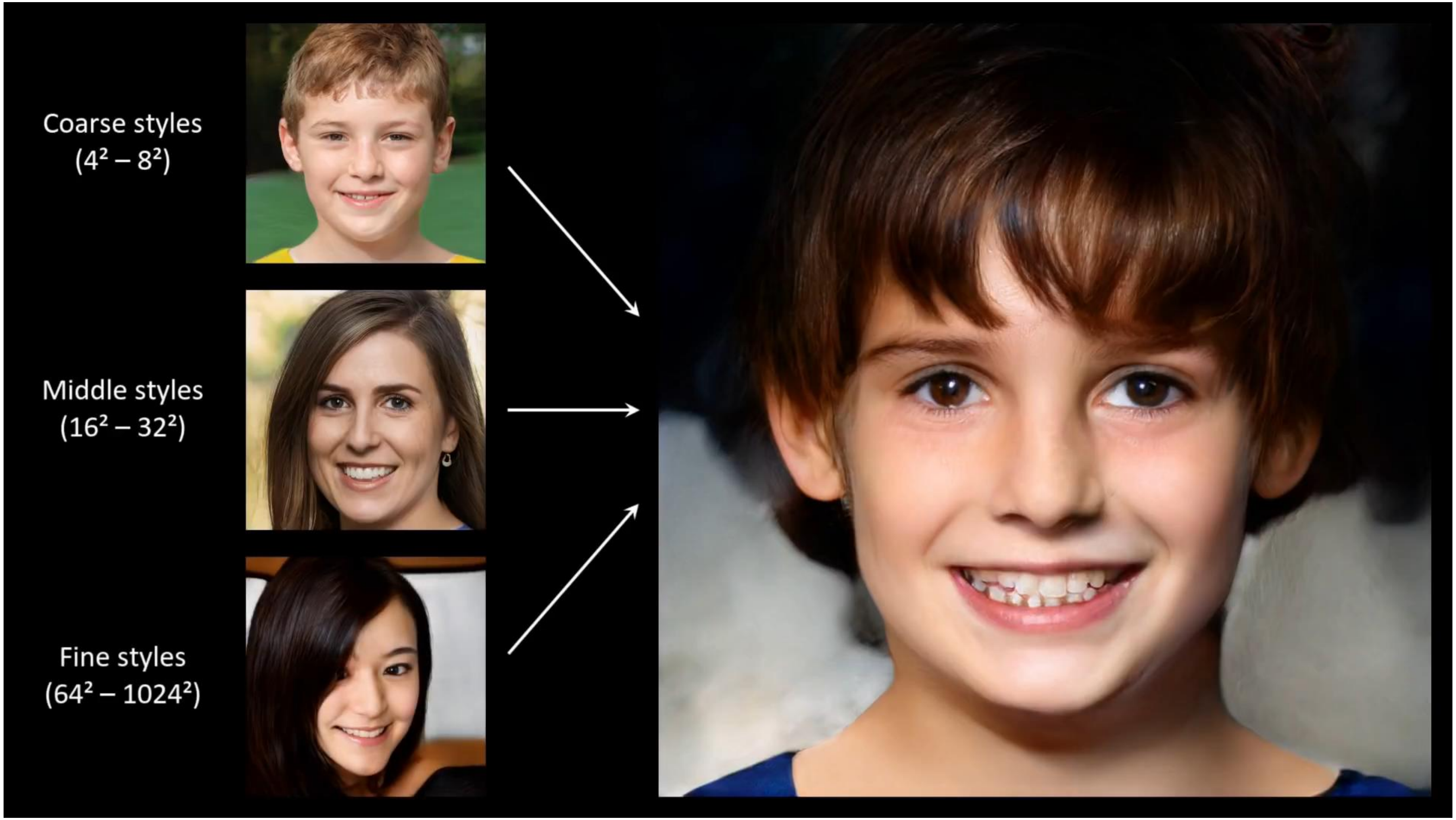
StyleGAN



https://youtu.be/kSLJriaOumA

[Karras et al. 19] StyleGAN

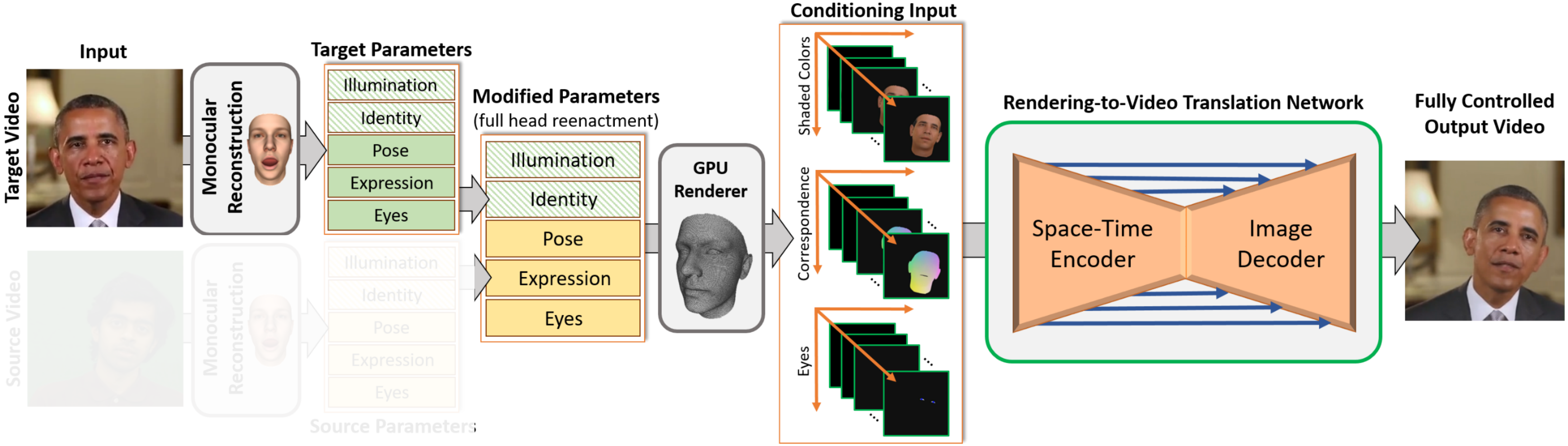
StyleGAN



https://youtu.be/kSLJriaOumA

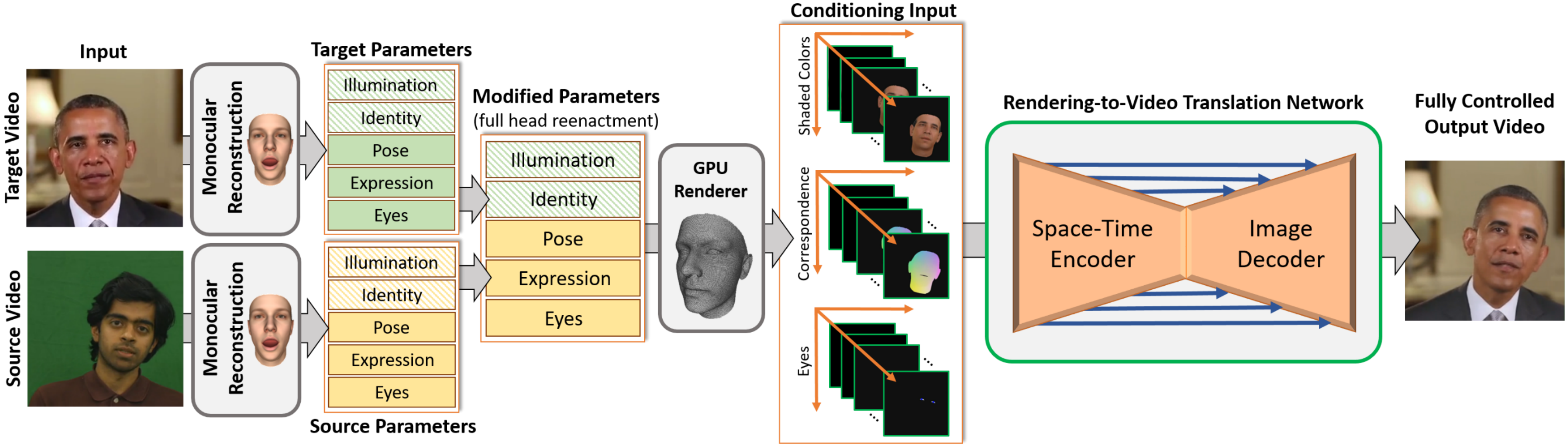
[Karras et al. 19] StyleGAN

Conditional GANs



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

Conditional GANs



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

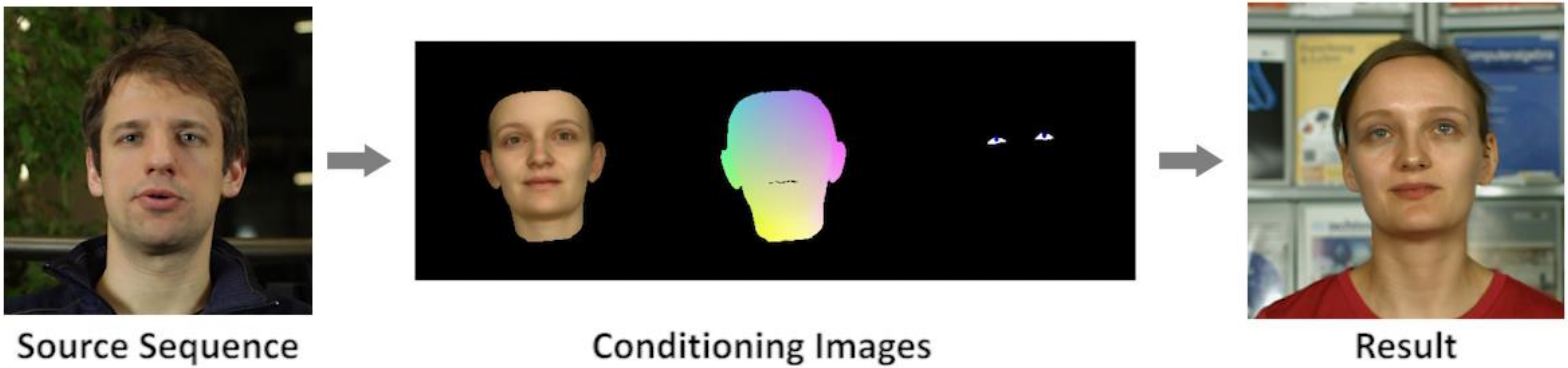
Conditioning on Face Reconstruction

Source Sequence | Conditioning Images | Result

Neural Network converts synthetic data to realistic video

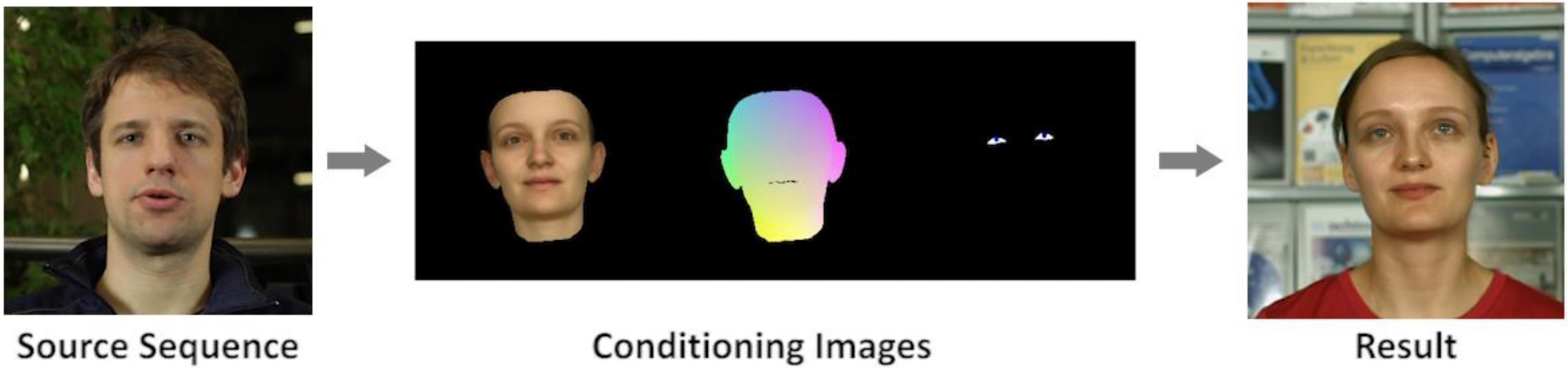
Siggraph’18 [Kim et al.]: Deep Video Portraits

Conditioning on Face Reconstruction



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

Conditioning on Face Reconstruction



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

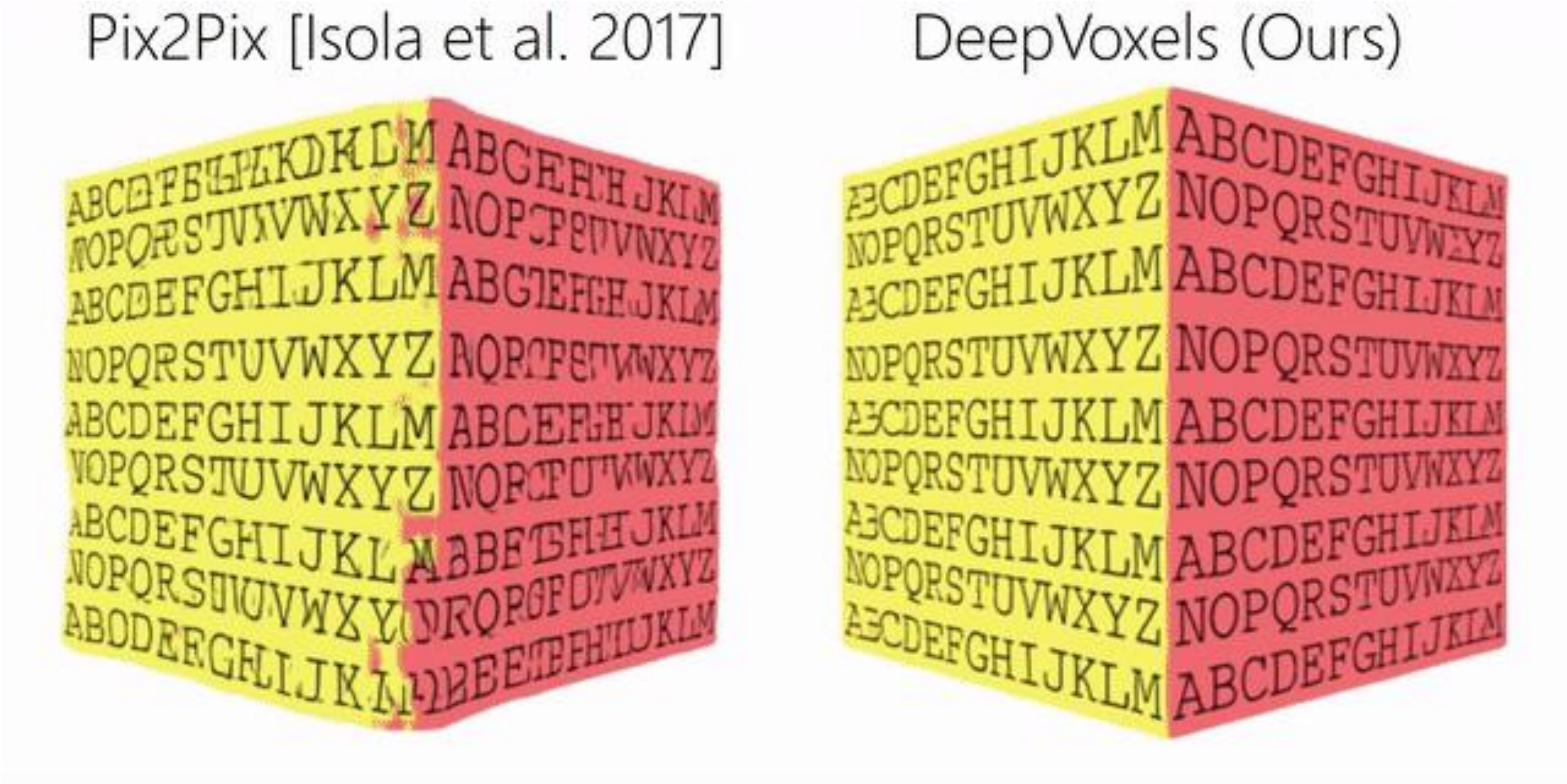
Video Editing

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

Videos still challenging for cGANs…

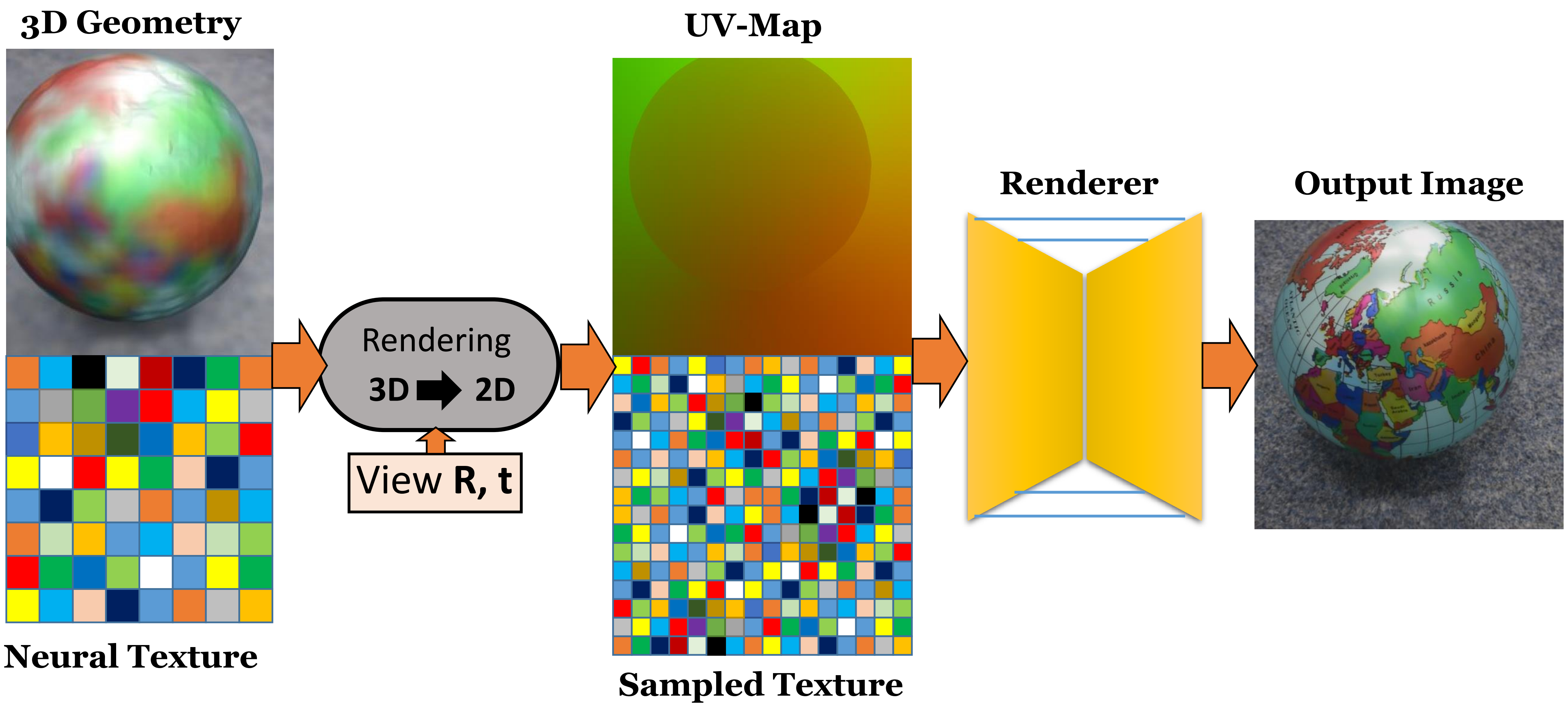
Pix2Pix [Isola et al. 2017]

DeepVoxels: Explicit 3D Features



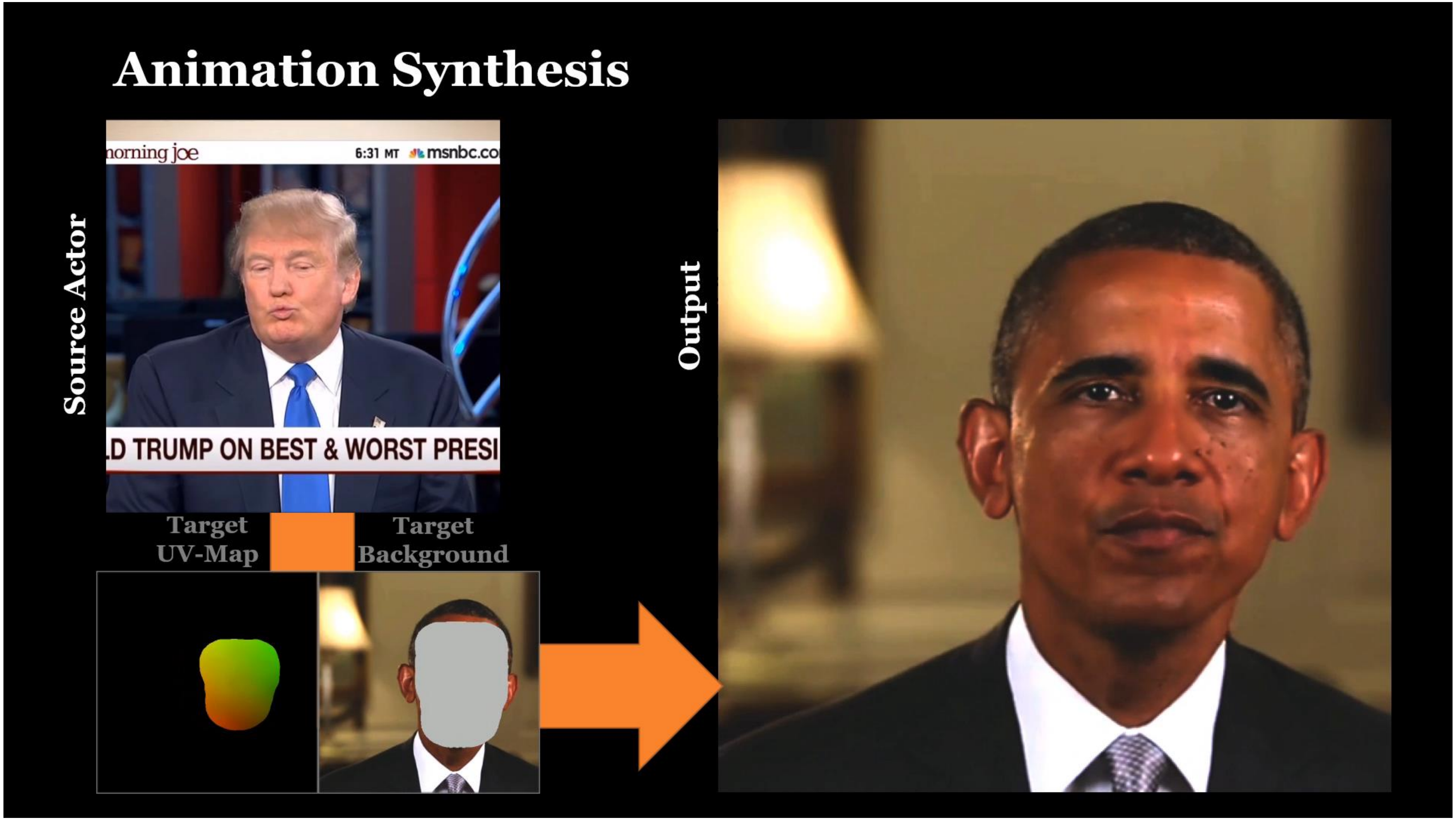
CVPR'19 (Oral) [Sitzmann et al.]: DeepVoxels

Neural Textures: Features on 3D Mesh



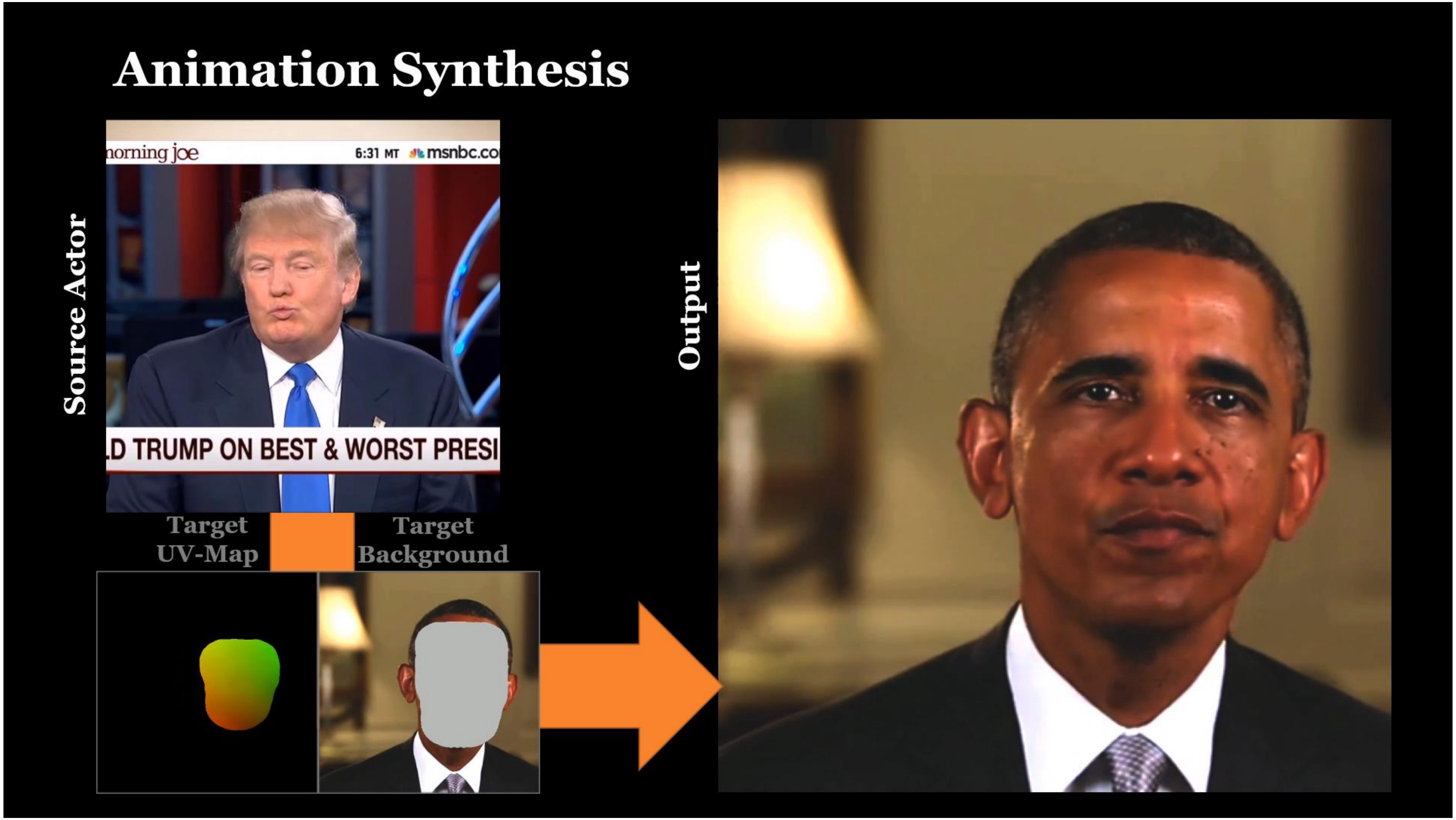
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



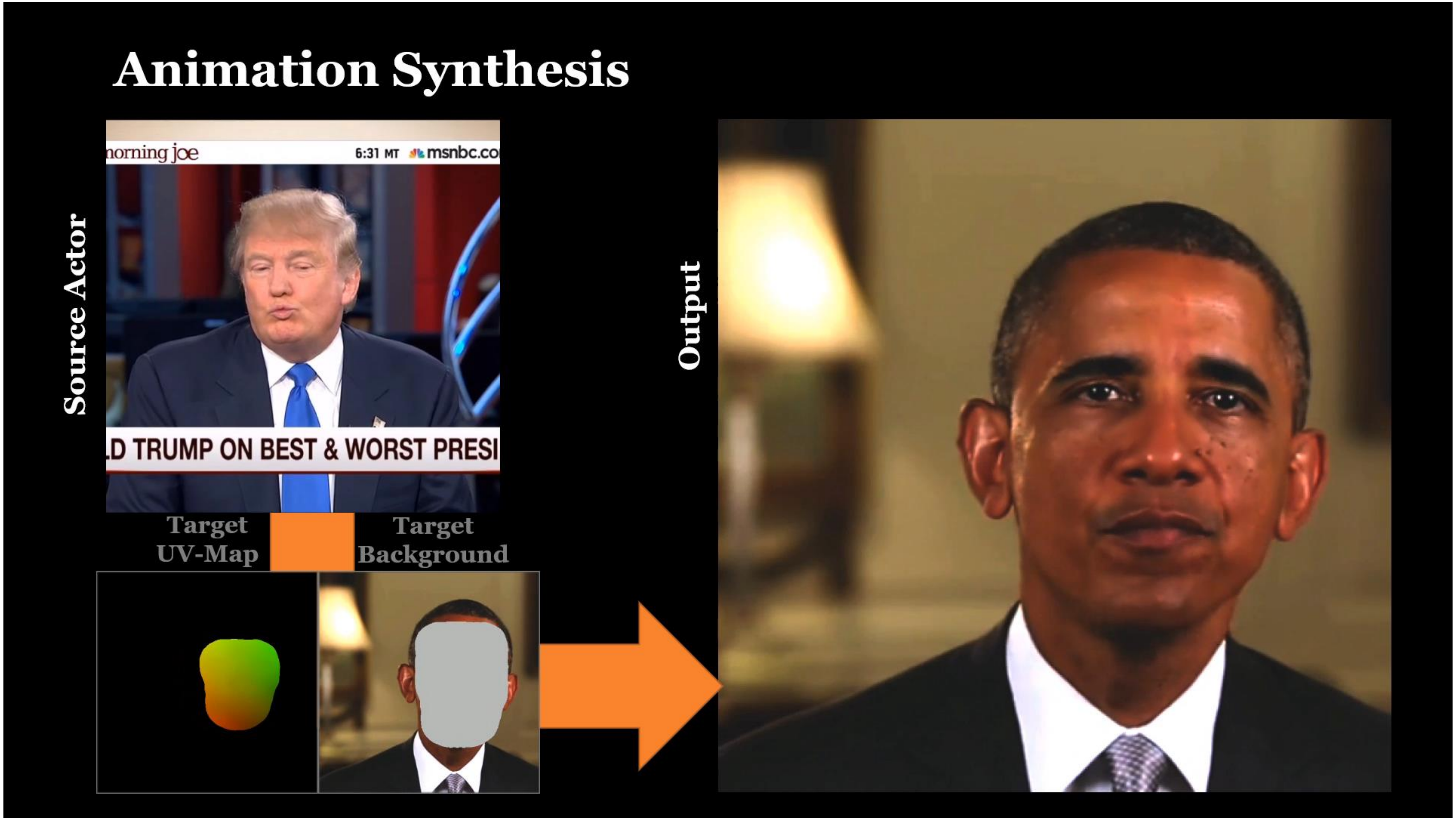
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



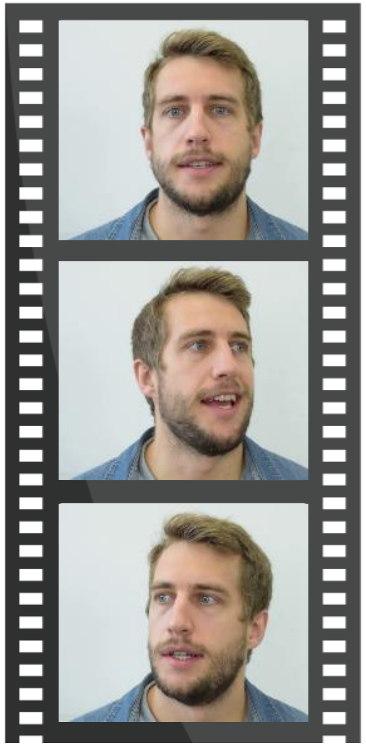
Siggraph’19 [Thies et al.]: Neural Textures

Facial Animation

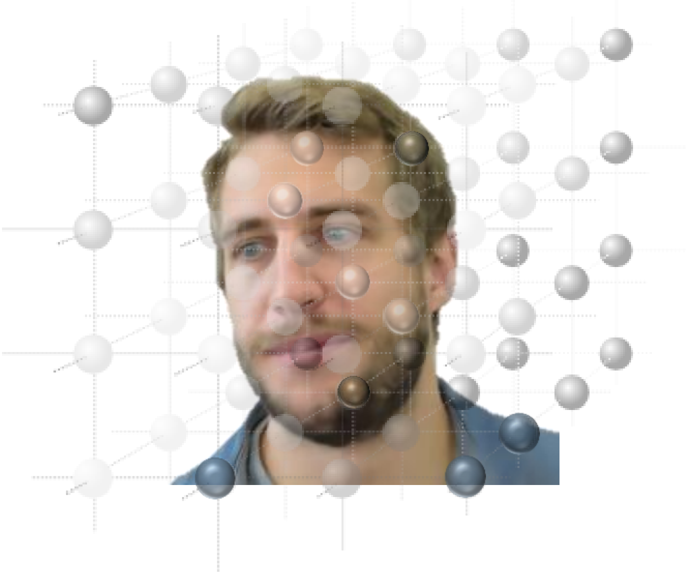


Siggraph'19 [Thies et al.]: Neural Textures

Dynamic Neural Radiance Fields for 4D Avatars



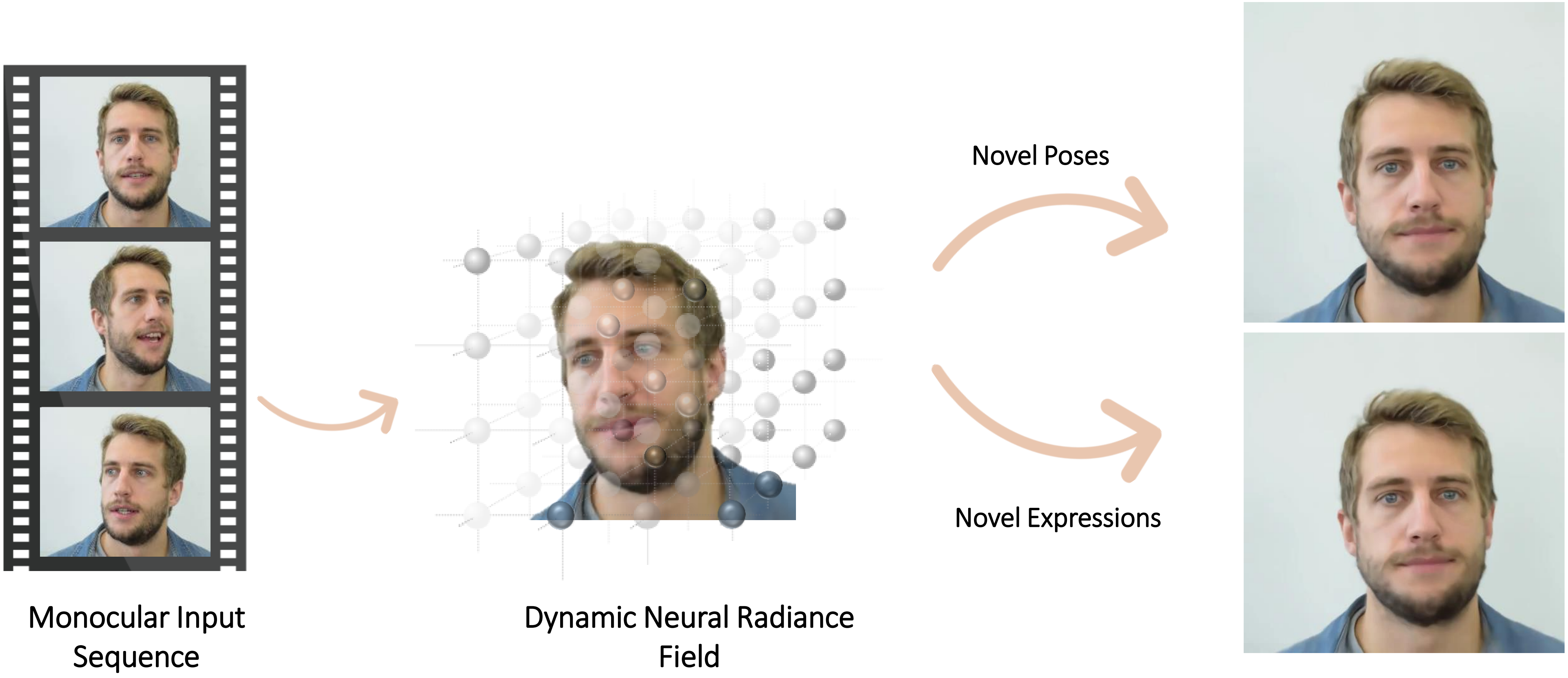
Monocular Input Sequence



Dynamic Neural Radiance Field

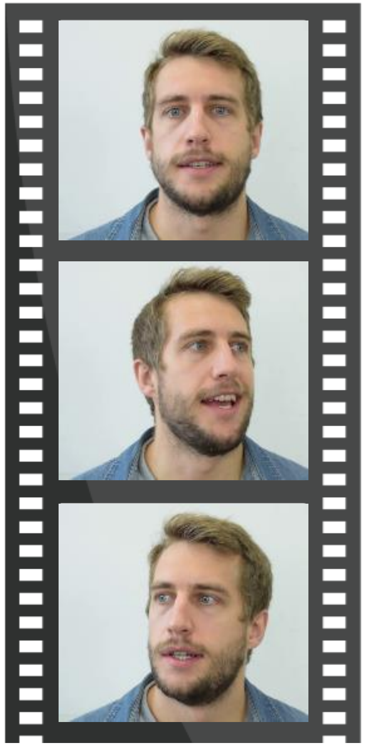
CVPR'21 [Gafni et al.]: Dynamic Neural Radiance Fields

Dynamic Neural Radiance Fields for 4D Avatars

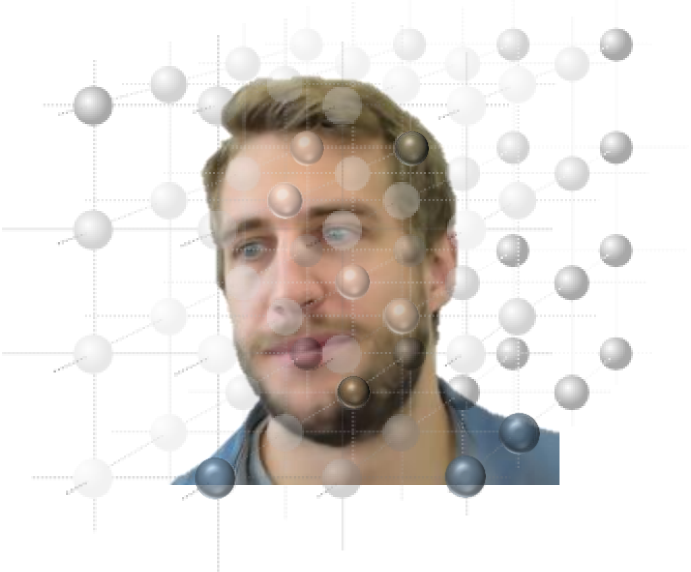


CVPR’21 [Gafni et al.]: Dynamic Neural Radiance Fields

Dynamic Neural Radiance Fields for 4D Avatars



Monocular Input Sequence



Dynamic Neural Radiance Field

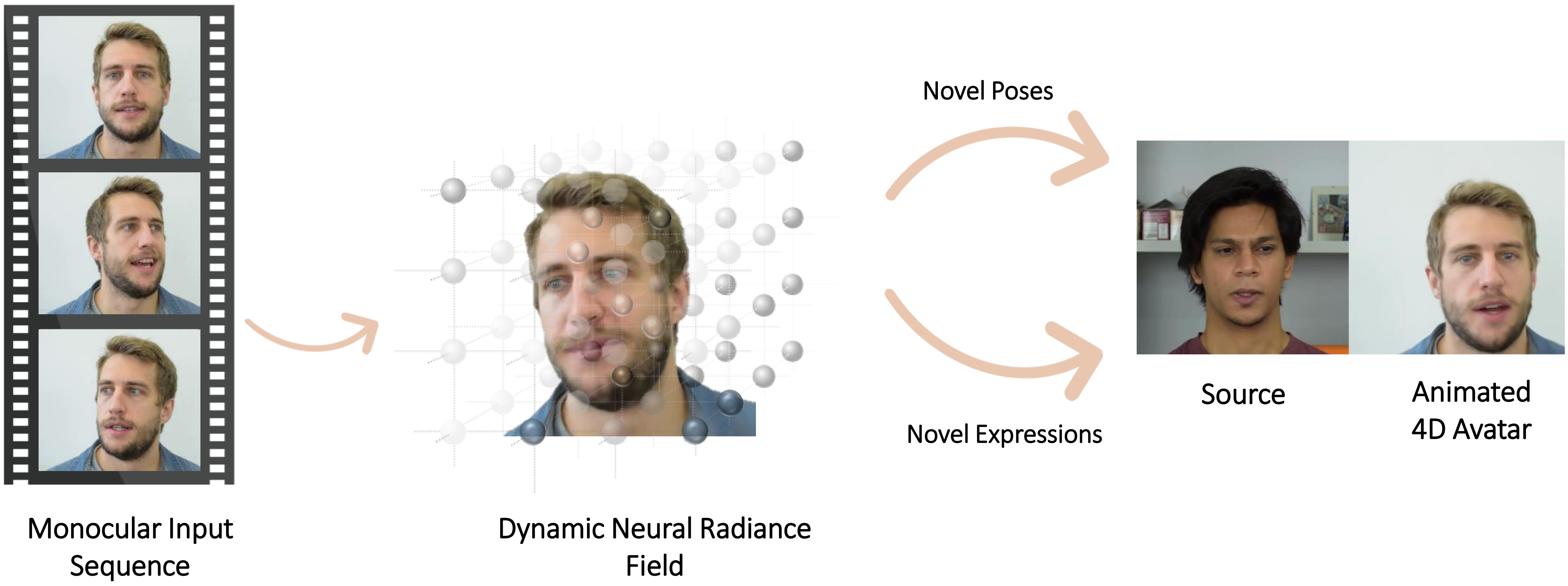
Novel Poses

Novel Expressions

4D Facial Avatar

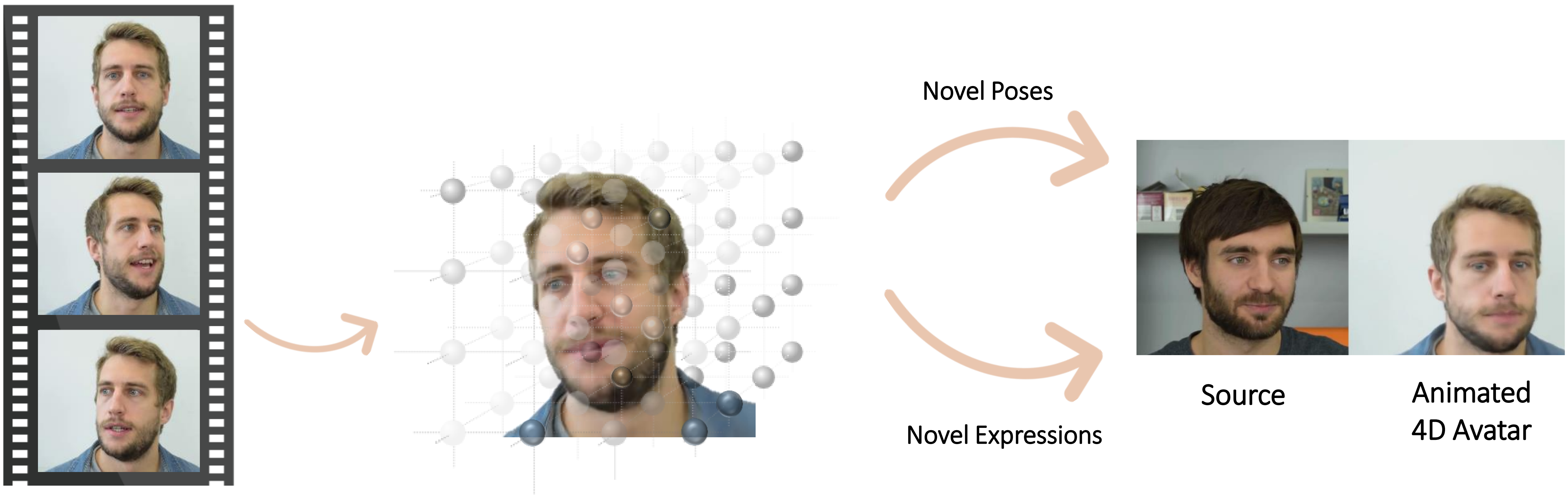
CVPR'21 [Gafni et al.]: Dynamic Neural Radiance Fields

Dynamic Neural Radiance Fields for 4D Avatars



CVPR'21 [Gafni et al.]: Dynamic Neural Radiance Fields

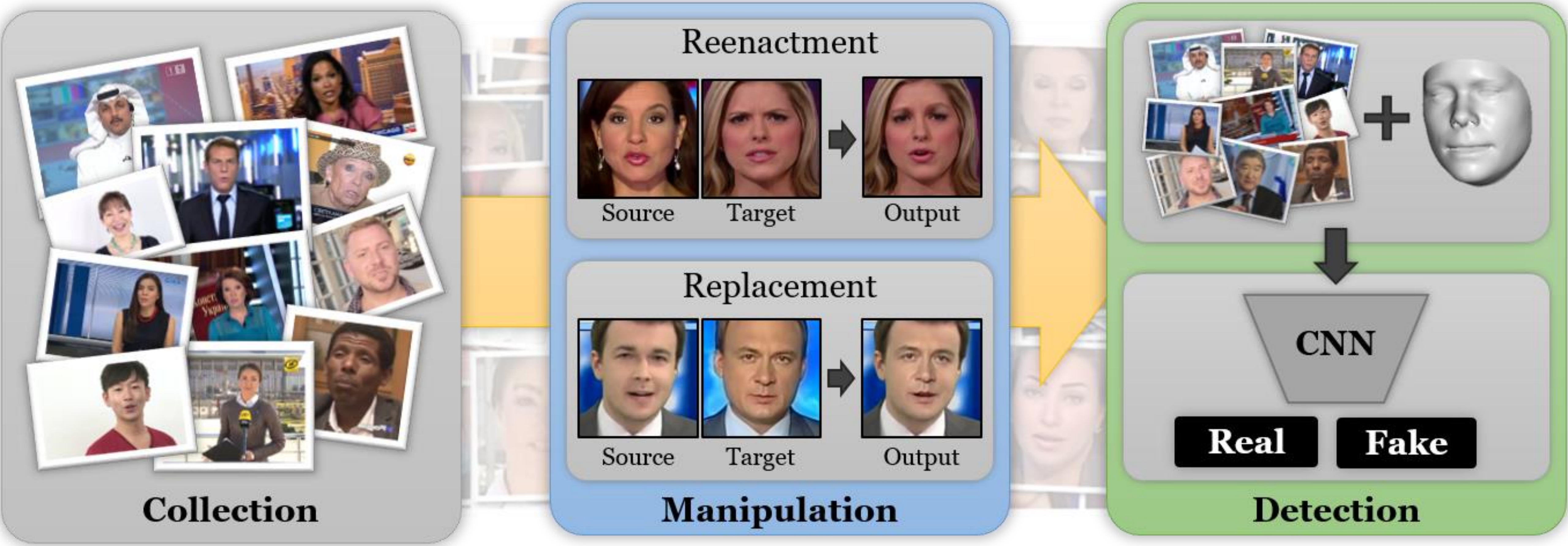
Dynamic Neural Radiance Fields for 4D Avatars



CVPR’21 [Gafni et al.]: Dynamic Neural Radiance Fields

What about Deep Fake Detection?

FaceForensics

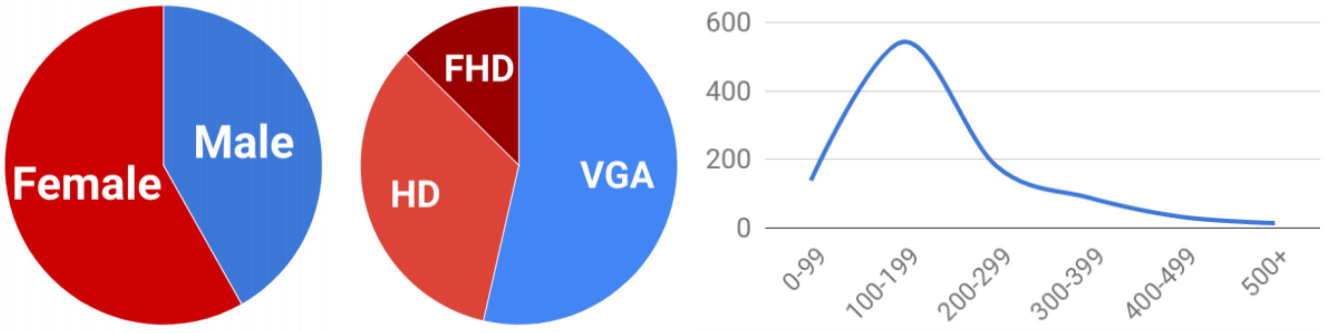


ICCV’19 [Roessler et al.]: FaceForensics

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 1000 research groups

ICCV'19 [Roessler et al.]: FaceForensics

FaceForensics: Benchmark

Method	Info	Deepfakes	Face2Face	FaceSwap	NeuralTextures	Pristine	Total
Xception	P	**0.964**	**0.869**	**0.903**	**0.807**	0.524	**0.710**
Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2019							
MesoNet		0.873	0.562	0.612	0.407	**0.726**	0.660
Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen: Mesonet: a compact facial video forgery detection network. arXiv							
XceptionNet Full Image	P	0.745	0.759	0.709	0.733	0.510	0.624
Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, Matthias Nießner: FaceForensics++: Learning to Detect Manipulated Facial Images. ICCV 2019							
Bayar and Stamm		0.845	0.737	0.825	0.707	0.462	0.616
Belhassen Bayar and Matthew C. Stamm: A deep learning approach to universal image manipulation detection using a new convolutional layer. ACM Workshop on Information Hiding and Multimedia Security							
Rahmouni		0.855	0.642	0.563	0.607	0.500	0.581
Nicolas Rahmouni, Vincent Nozick, Junichi Yamagishi, and Isao Echizen: Distinguishing computer graphics from natural images using convolution neural networks. IEEE Workshop on Information Forensics and Security,							
Recasting		0.855	0.679	0.738	0.780	0.344	0.552
Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. ACM Workshop on Information Hiding and Multimedia Security							
Steganalysis Features		0.736	0.737	0.689	0.633	0.340	0.518
Jessica Fridrich and Jan Kodovsky: Rich Models for Steganalysis of Digital Images. IEEE Transactions on Information Forensics and Security							

On 700 high-quality images + hidden test set + automated evaluation

ICCV'19 [Roessler et al.]: FaceForensics

Unsupervised / Self-Supervised Forensics

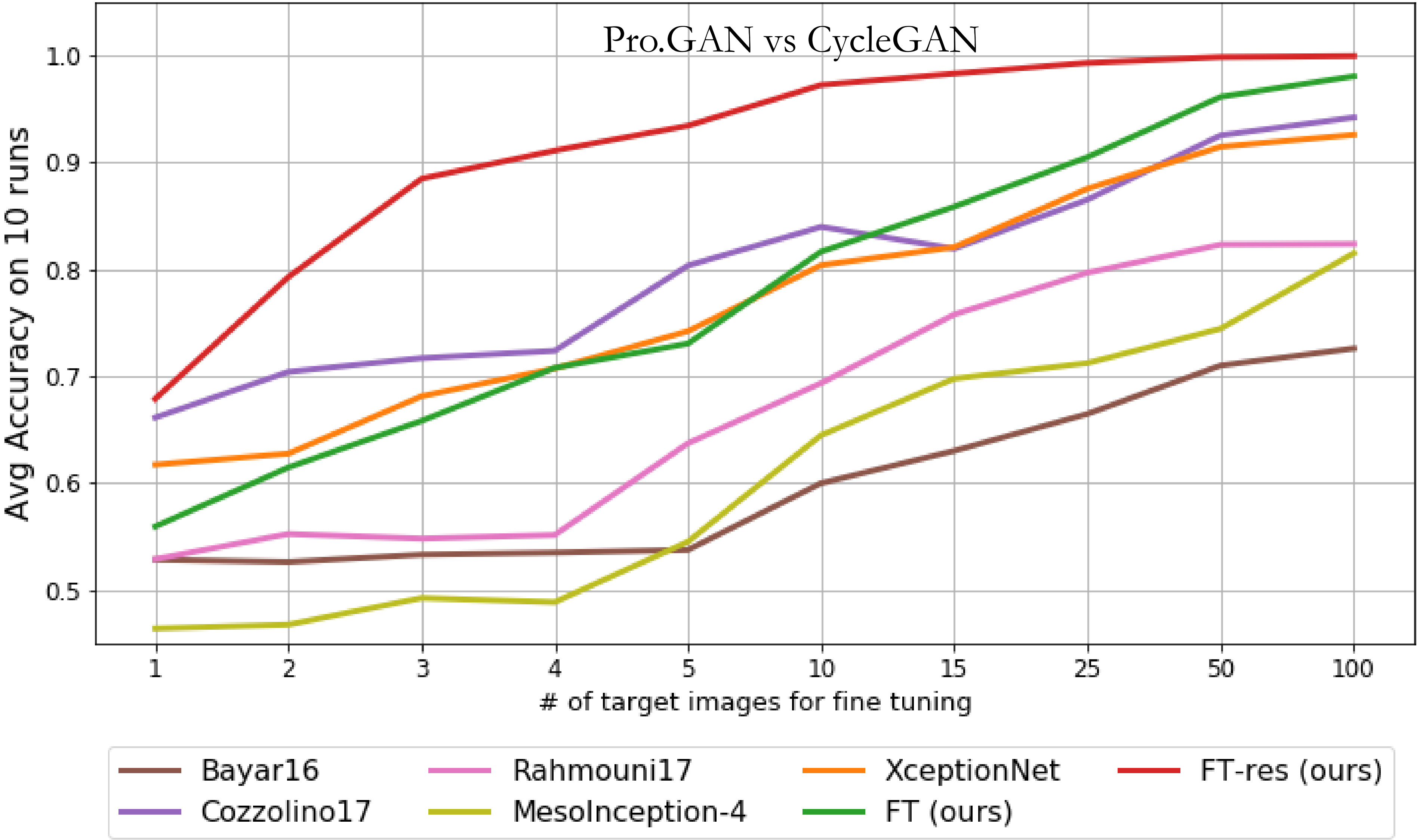
Major challenges

- Self-supervised Learning
- Transfer Learning
- Unsupervised Learning

XceptionNet	Test on Face2Face	Test on FaceSwap
Trained on Face2Face	98.13%	50.20%
Trained on FaceSwap	52.73%	98.30%

[Cozzolino et al. 20]: ForensicTransfer

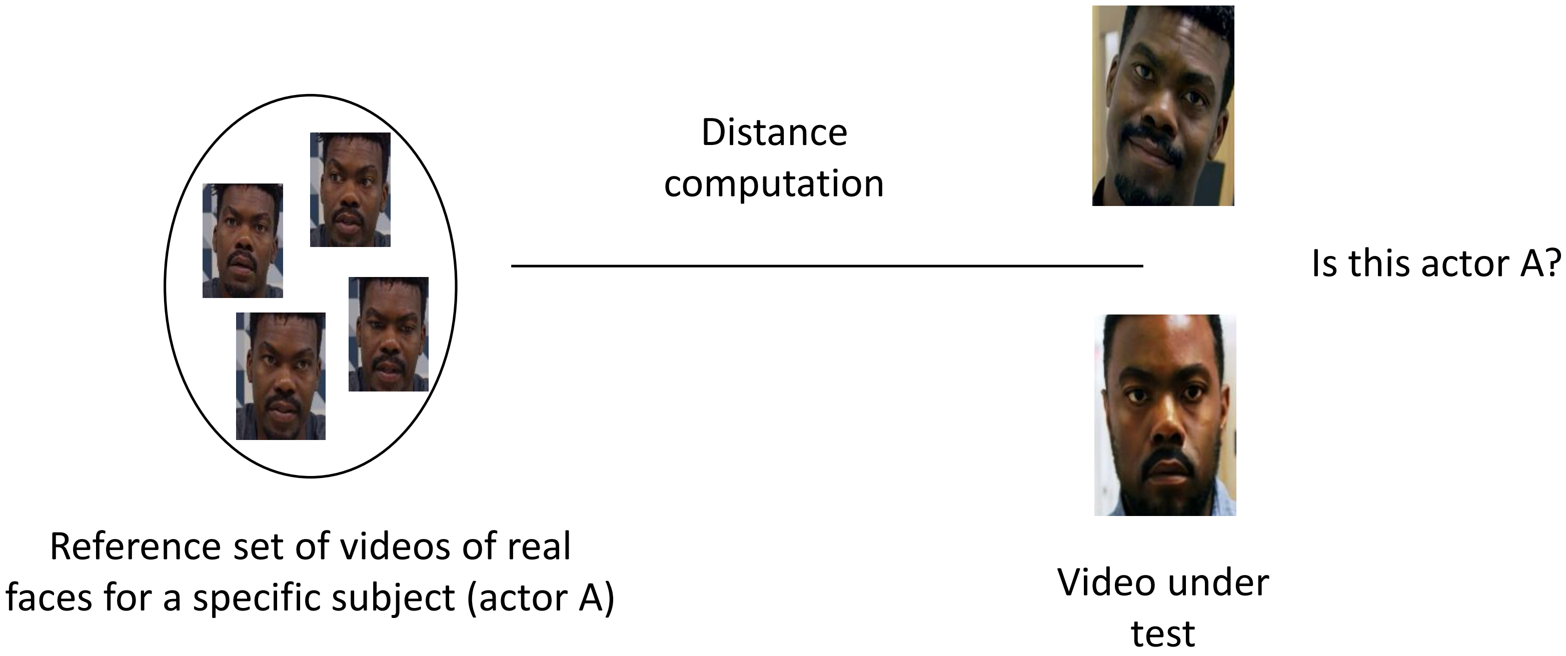
Forensic Transfer: Few Shot Learning



[Cozzolino et al. 20]: ForensicTransfer

ID-Reveal

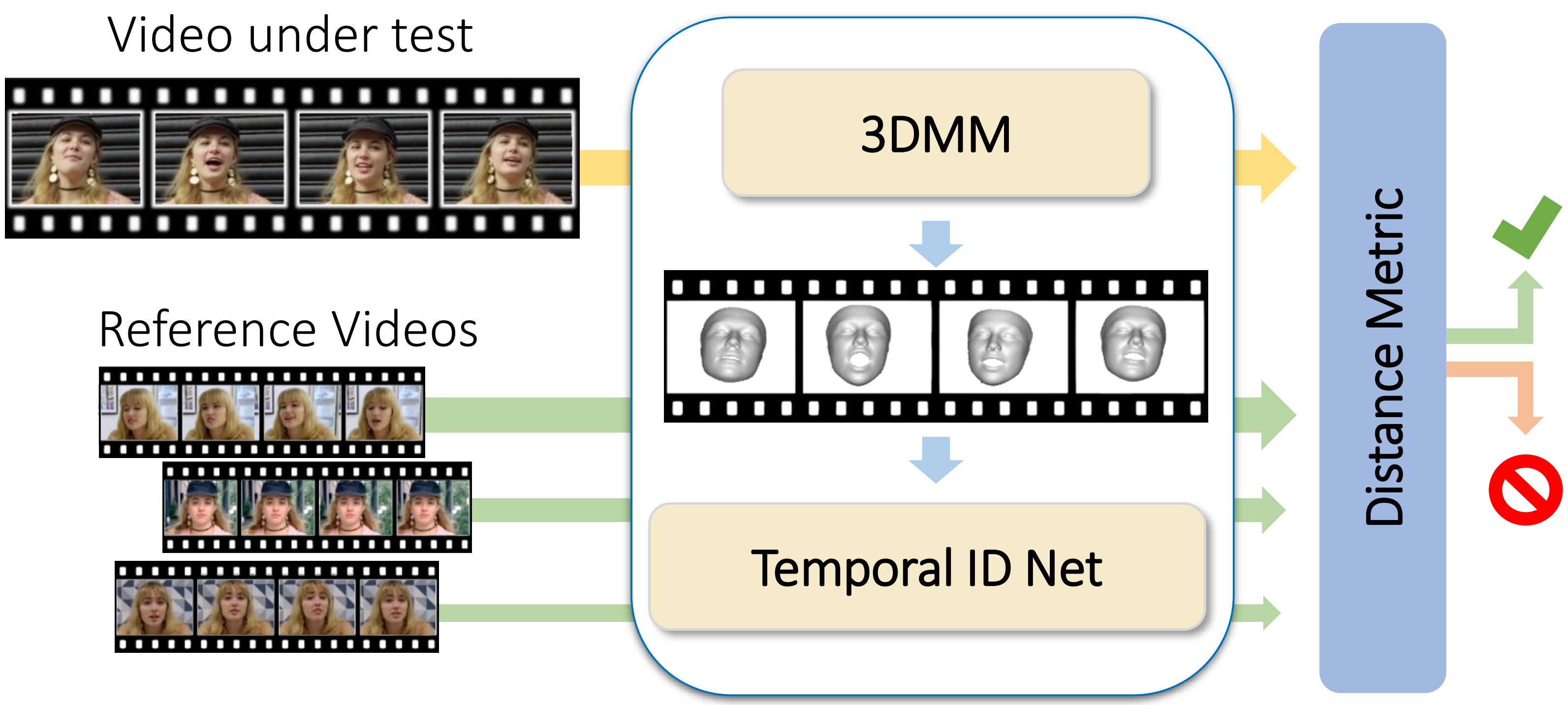
- It is trained only on real videos (Voxceleb, more than 5000 identities)
- It captures the biometrics of a specific identity
- Is this the identity of that specific subject?



[Cozzolino et al. 20]: ID-Reveal

Proposed Approach

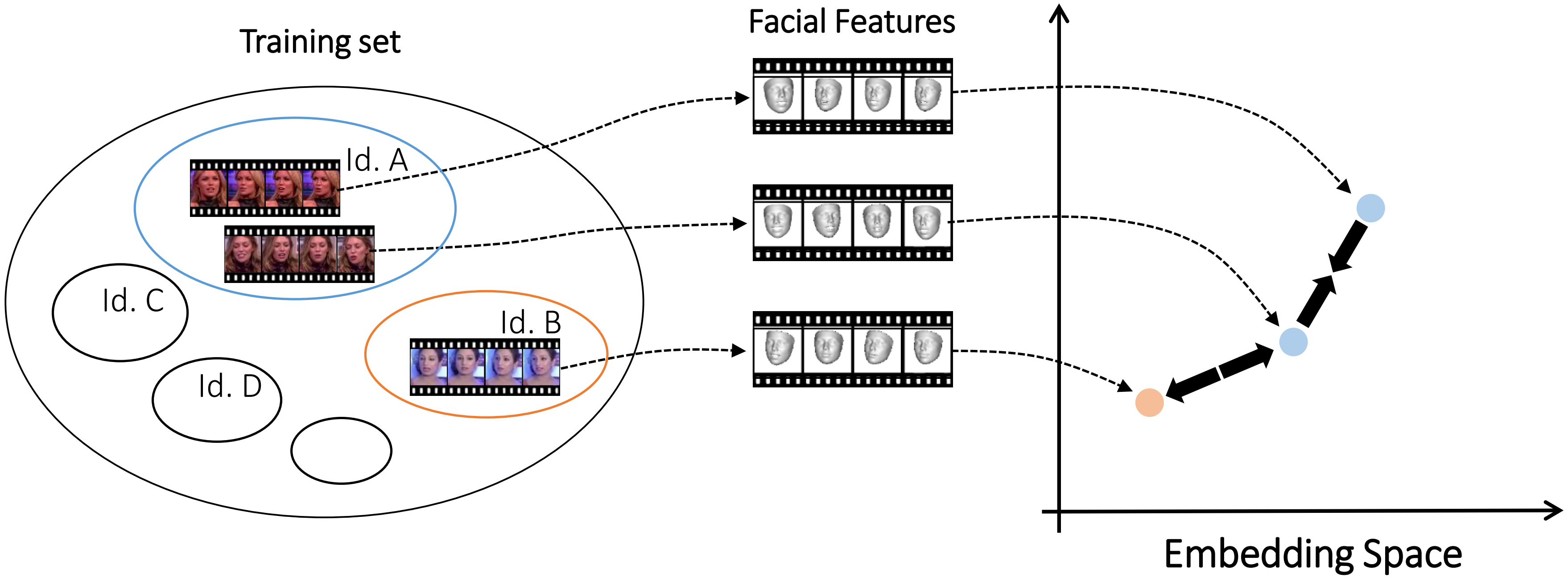
- Spatio-temporal feature extraction + adversarial training



[Cozzolino et al. 20]: ID-Reveal

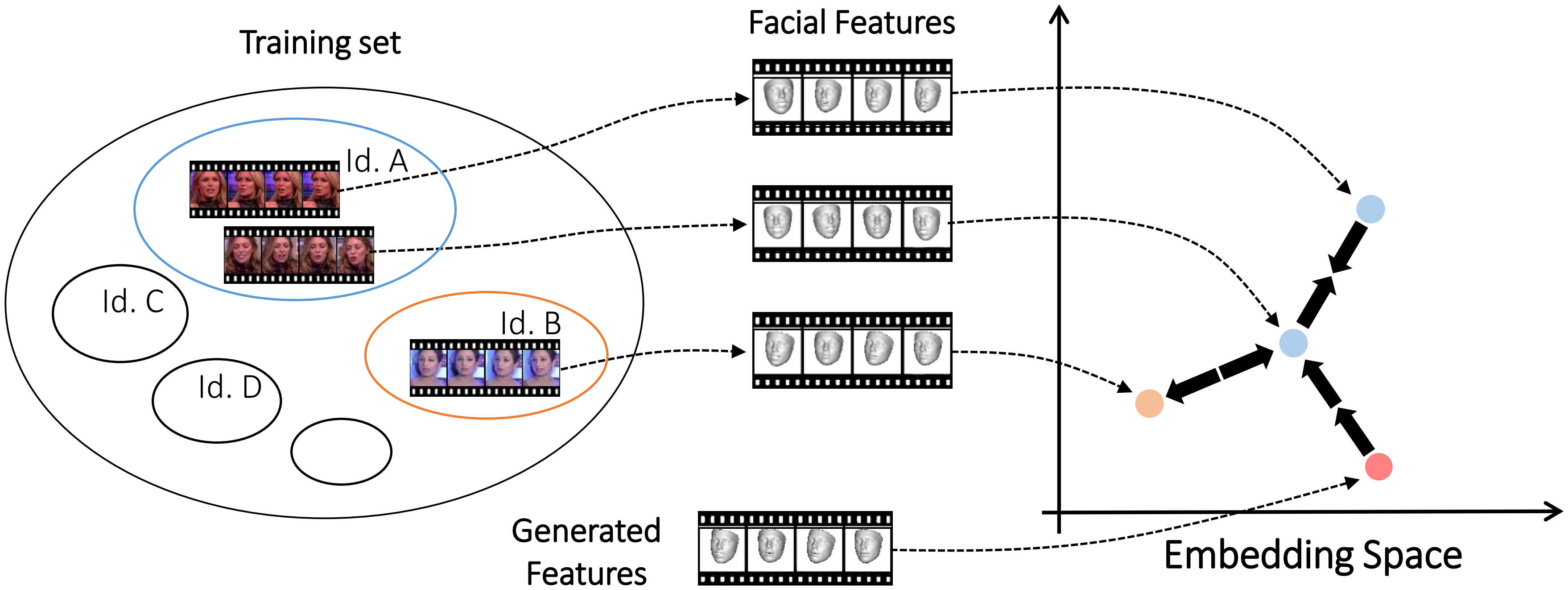
Metric Learning

- For each face we extract features (shape, expression, pose) obtained using the 3D morphable model
- The network is trained so as that the embedded vectors of the same subject are close but far from those of different subjects

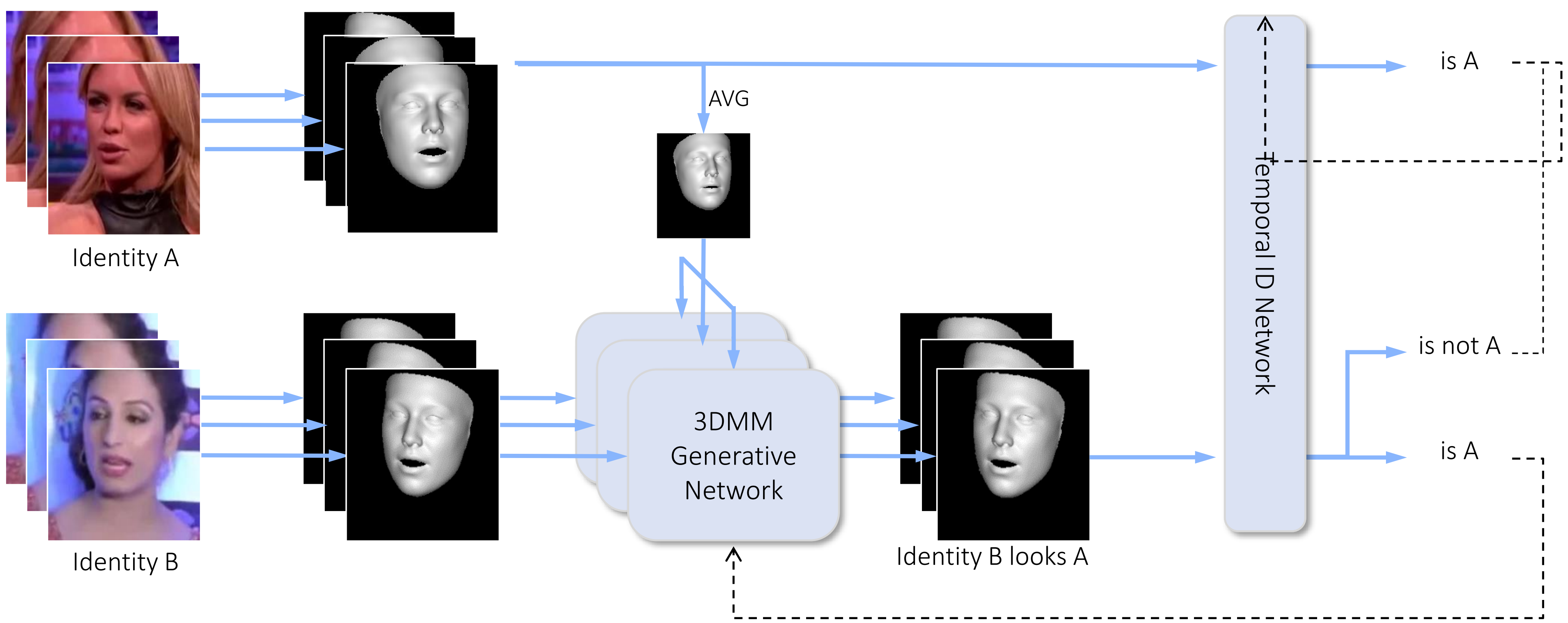


Adversarial Training

- We use a generative network to produce features similar to those we may extract from a manipulated video
- The objective of the adversarial game is to increase the ability of the network to distinguish real identities from fake ones

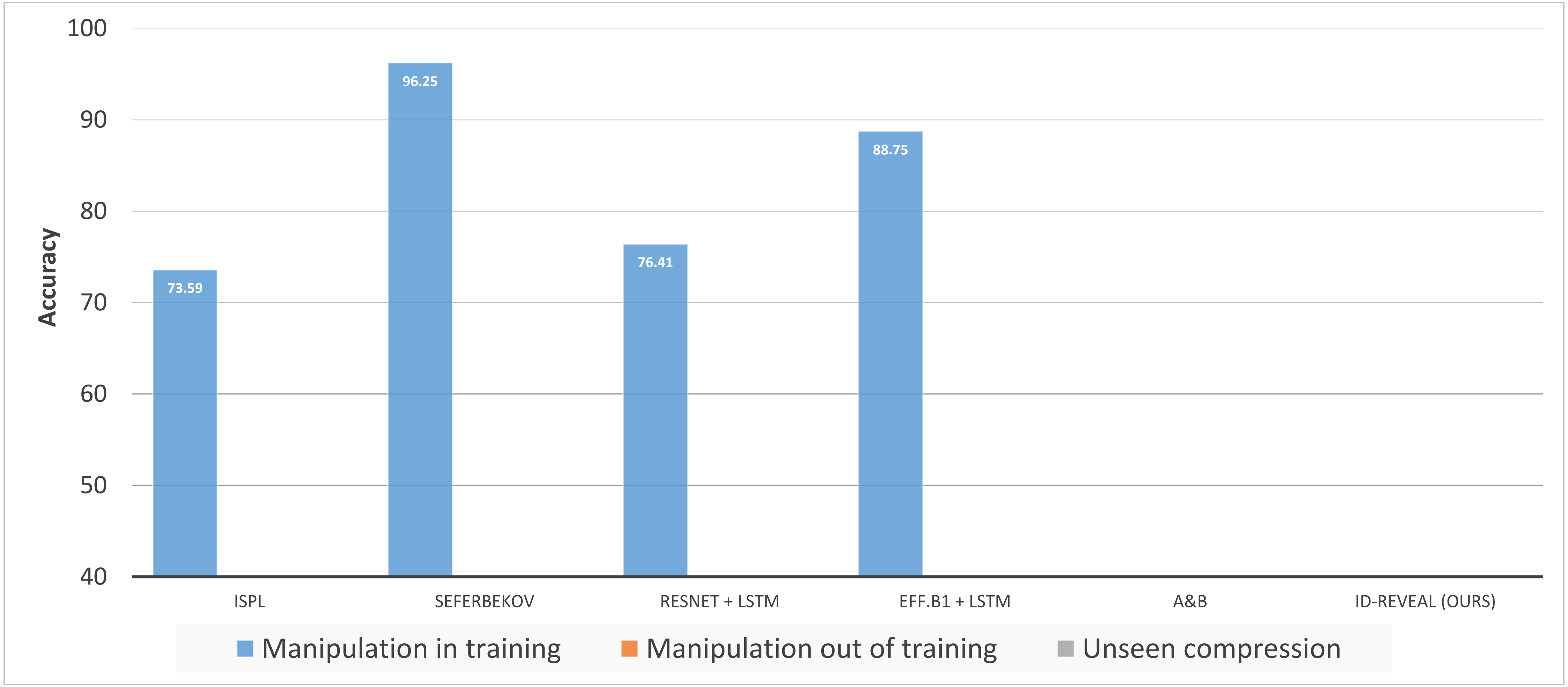


Training

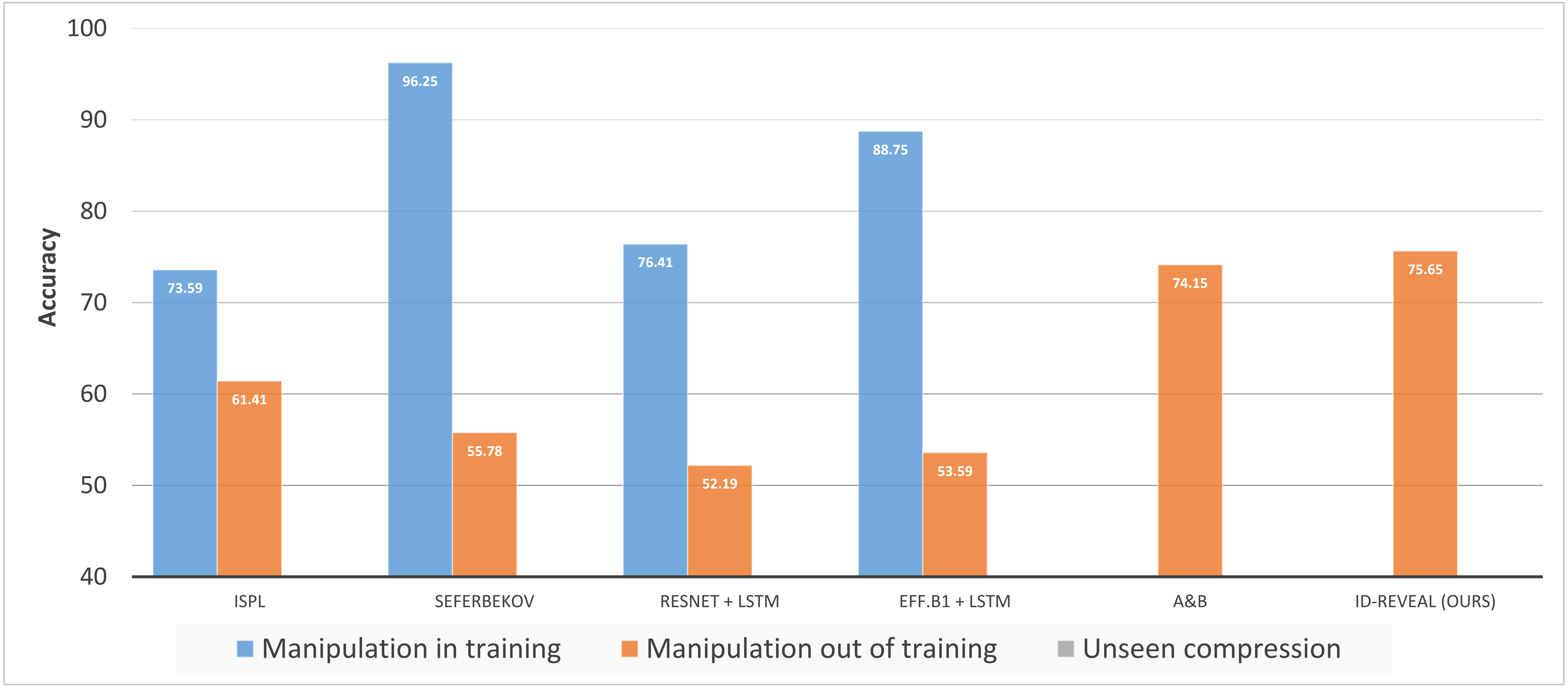


Identity A
Identity B
AVG
3DMM
Generative
Network
Identity B looks A
Temporal ID Network
is A
is not A
is A

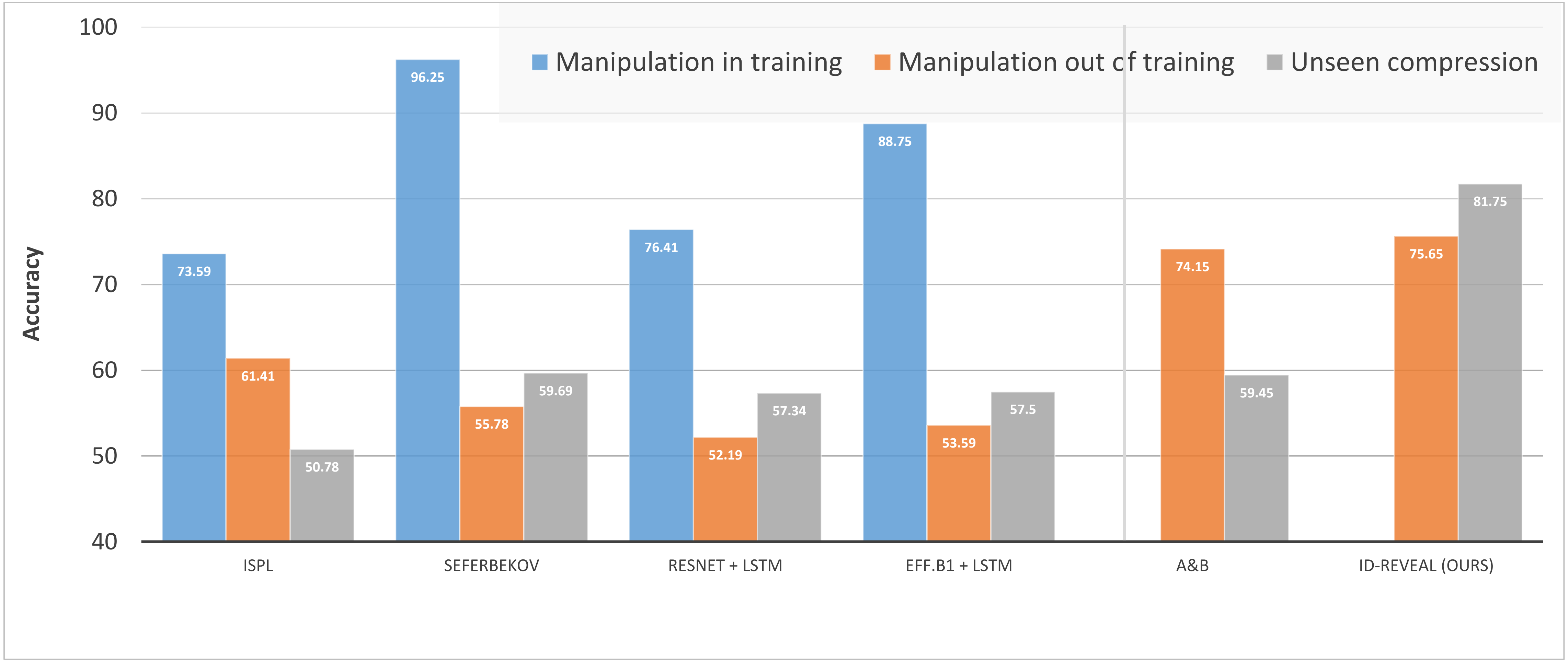
Detection Results



Detection Results

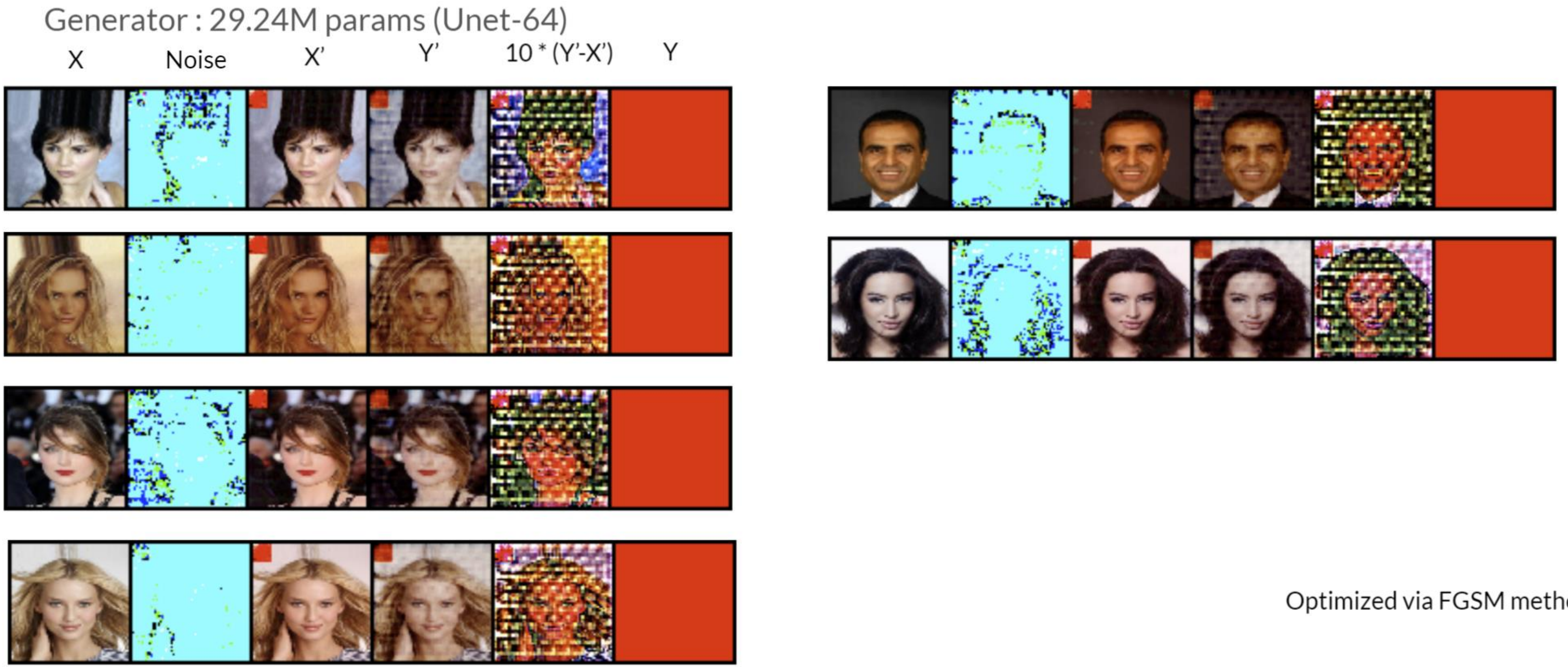


Detection Results



Some Work in Progress

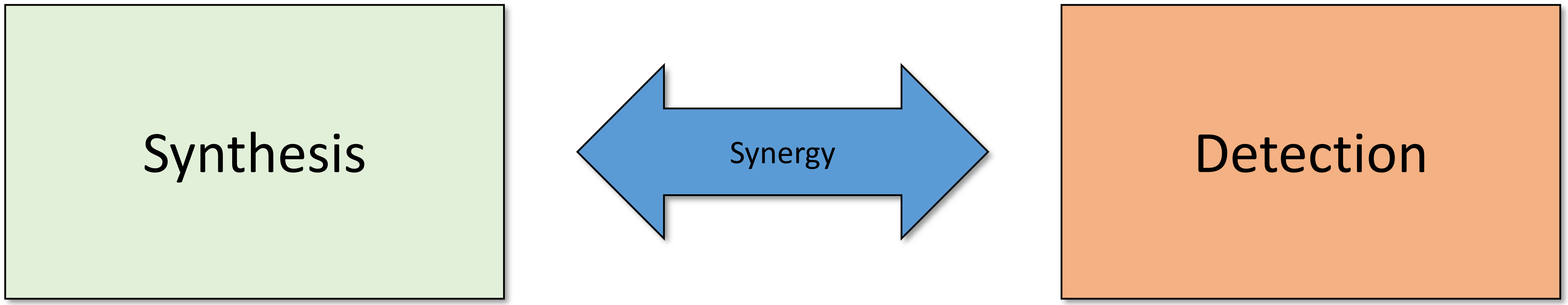
Active Defense against Generative Models?



Optimized via FGSM method

$$\text{Loss} = L1(Y', Y) + || \Delta ||$$

Conclusion



Thank You!


Justus Thies

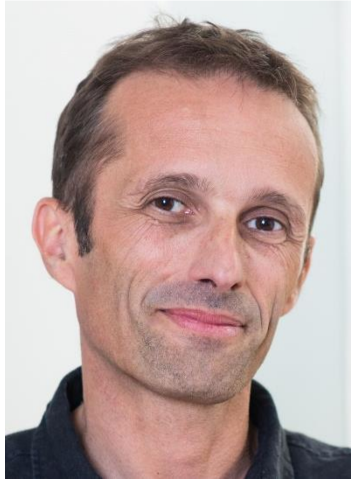

Michael Zollhöfer


Marc Stamminger


Christian Theobalt

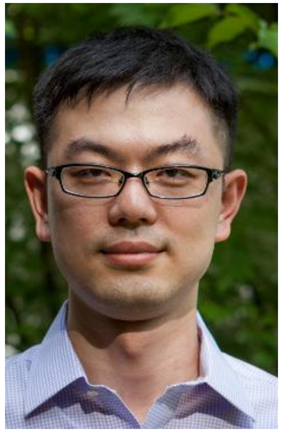

Christian Richardt


Christian Riess


Patrick Pérez


Andreas Rössler

Ayush Tewari


Weipeng Xu


Pablo Garrido


Levi Valgaerts


Hyeongwoo Kim


Davide Cozzolino


Luisa Verdoliva

Visual Computing & Artificial Intelligence
Prof. Matthias Nießner

CVPR VIRTUAL JUNE 19-25

Papers at a Glance

TUM VISUAL COMPUTING

Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction

Keywords: facial re-enactment, 4D reconstruction

Poster Q/A: 23rd June, paper session #7

Project Page

SPSG: Self-Supervised Photometric Scene Generation from RGB-D Scans

Keywords:

Poster Q/A: 21st June, paper session #2

Project Page

Seeing Behind Objects for 3D Multi-Object Tracking in RGB-D Sequences

Keywords: 3D reconstruction, tracking

Poster Q/A: 22nd June, paper session #5

Project Page

Neural Deformation Graphs for Globally-consistent Non-rigid Reconstruction

Keywords: non-rigid 3D reconstruction

Poster Q/A: 21st June, paper session #2

Project Page

RfD-Net: Point Scene Understanding by Semantic Instance Reconstruction

Keywords: 3D scene understanding, instance reconstruction

Poster Q/A: 22nd June, paper session #4

Project Page

Exploring Data-Efficient 3D Scene Understanding with Contrastive Scene Contexts

Keywords: data-efficient, 3D scene understanding

Poster Q/A: 25th June, paper session #12

Project Page

Scan2Cap: Context-aware Dense Captioning in RGB-D Scans

Keywords: dense captioning, natural language

Poster Q/A: 22nd June, paper session #3

Project Page