More Generative Models ☺

Conditional GANs on Videos

- Challenge:
  - Each frame is high quality, but temporally inconsistent

Video-to-Video Synthesis

- Sequential Generator: $$p(\tilde{\mathbf{x}}_1^T | \mathbf{s}_1^T) = \prod_{t=1}^{T} p(\tilde{\mathbf{x}}_t | \tilde{\mathbf{x}}_{t-L}^{t-1}, \mathbf{s}_{t-L}^{t}).$$

past L generated frames

past L source frames

(set L = 2)

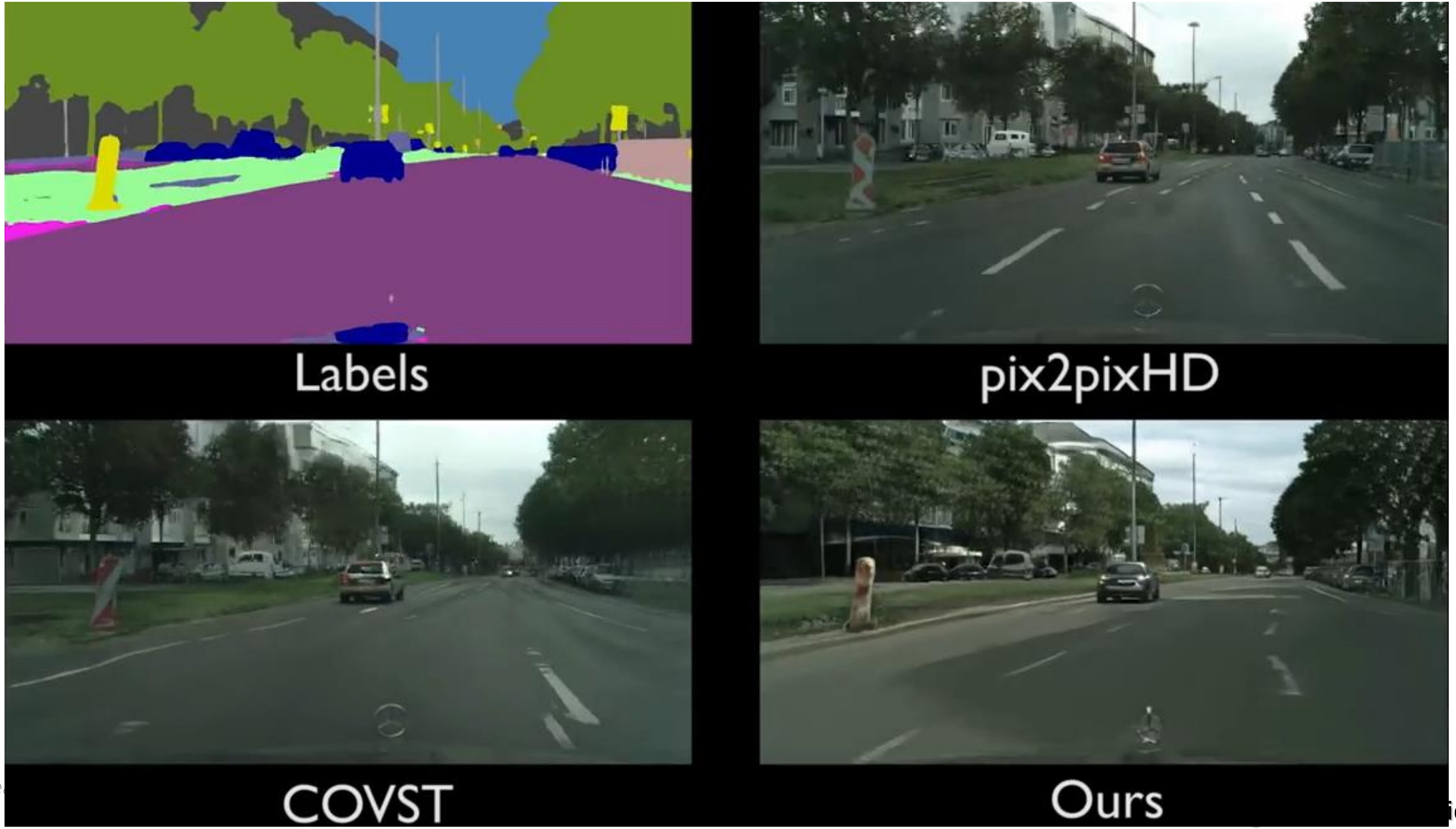
- Conditional Image Discriminator $\boldsymbol{D_i}$ (is it real image)
- Conditional Video Discriminator $\boldsymbol{D_v}$ (temp. consistency via flow)

Full Learning Objective:

$$\min_F \left(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V) \right) + \lambda_W \mathcal{L}_W(F),$$

Wang et al. 18: Vid2Vid

Video-to-Video Synthesis



Video-to-Video Synthesis

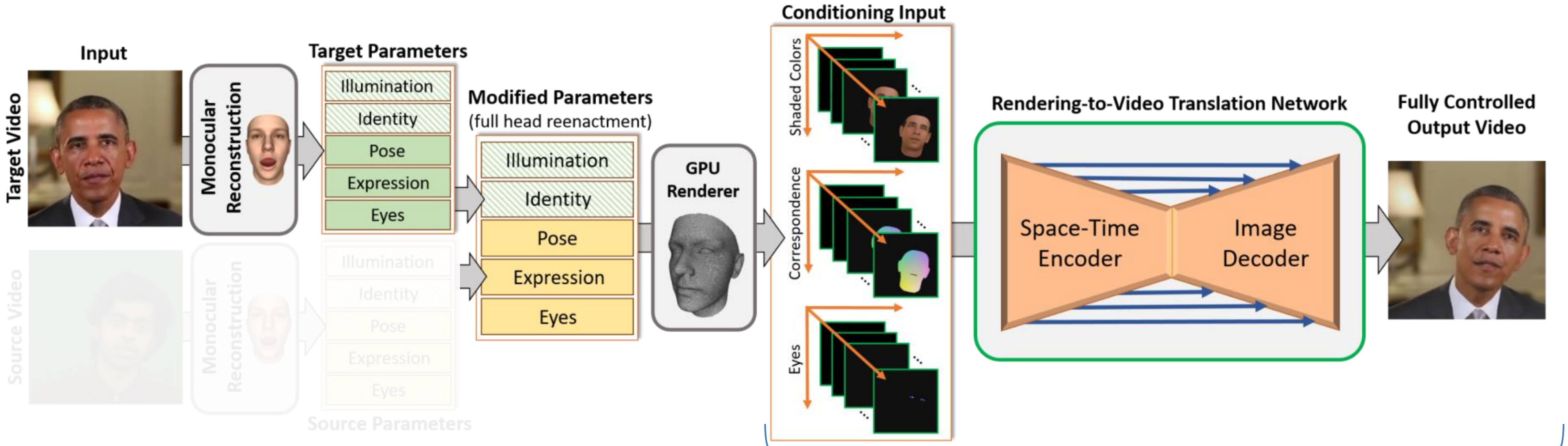
- Key ideas:
  - Separate discriminator for temporal parts
    - In this case based on optical flow
  - Consider recent history of prev. frames
  - Train all of it jointly

Wang et al. 18: Vid2Vid

Deep Video Portraits

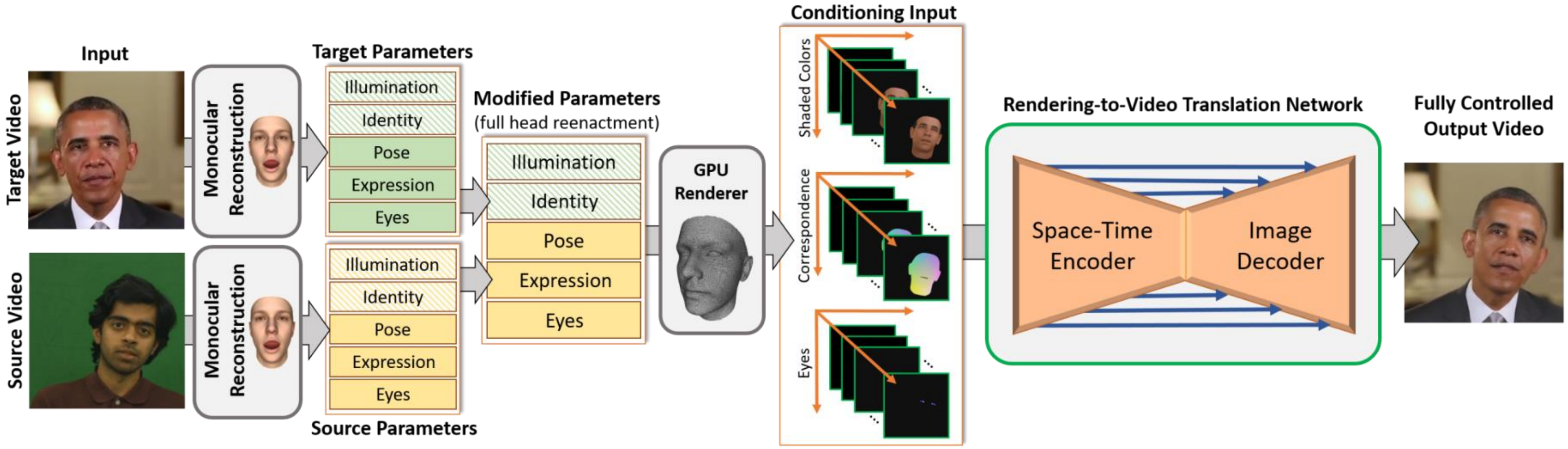
Siggraph'18 [Kim et al 18]: Deep Portraits

Deep Video Portraits



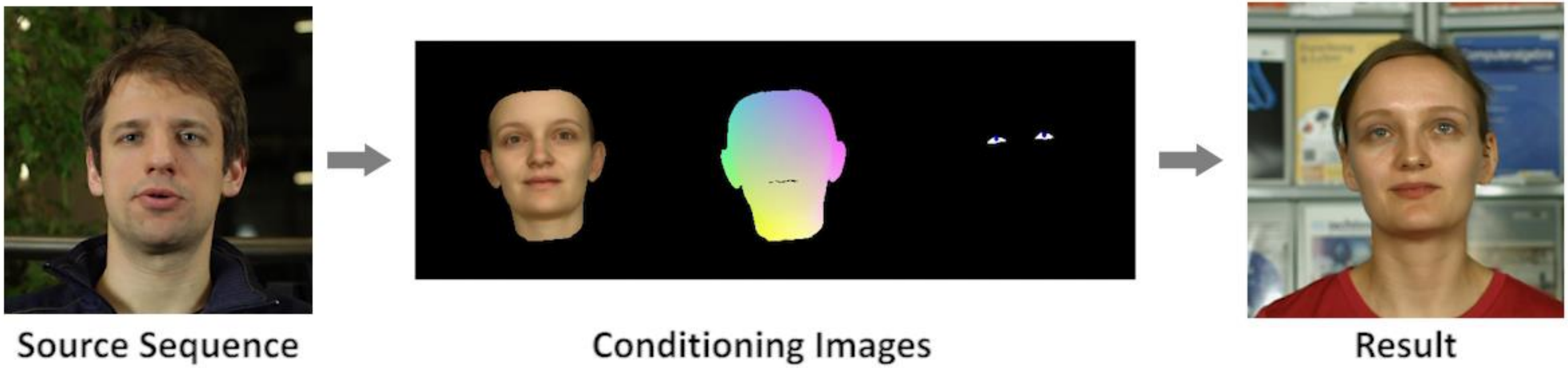
Similar to “Image-to-Image Translation” (Pix2Pix) [Isola et al.]

Deep Video Portraits



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

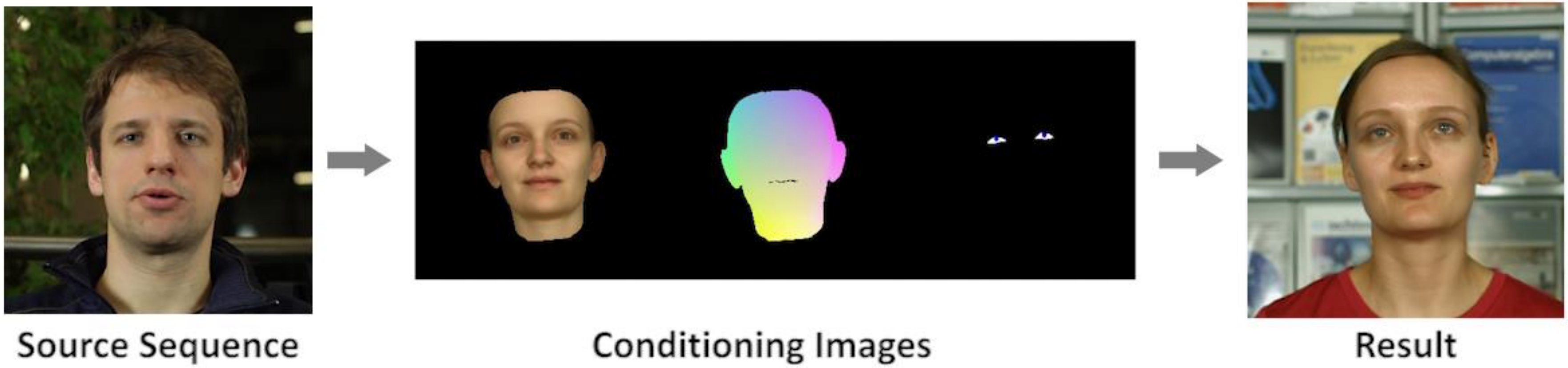
Deep Video Portraits



Neural Network converts synthetic data to realistic video

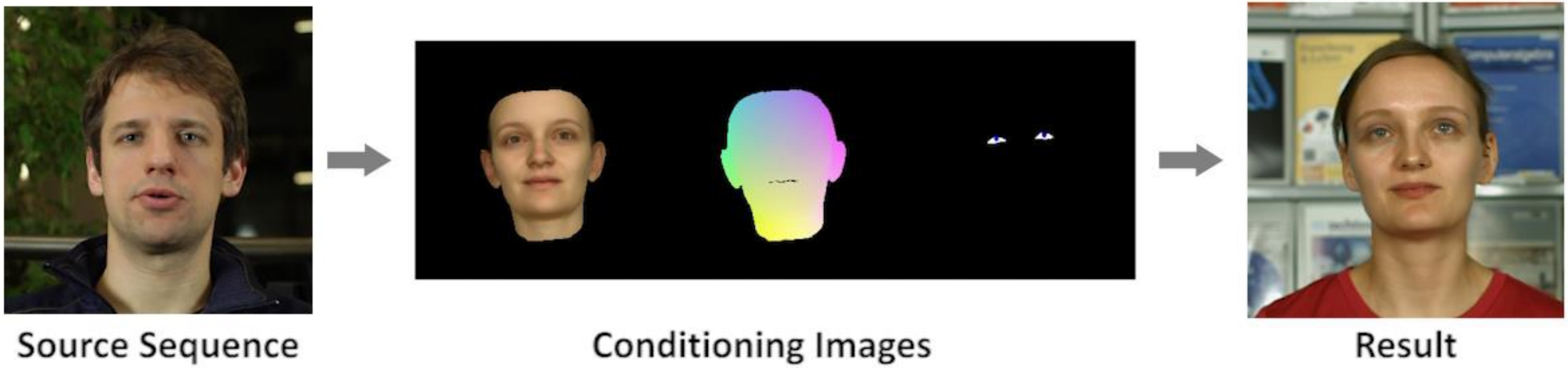
Siggraph'18 [Kim et al 18]: Deep Portraits

Deep Video Portraits



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

Deep Video Portraits



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

Deep Video Portraits

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

Deep Video Portraits

Interactive Video Editing

2x speed

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

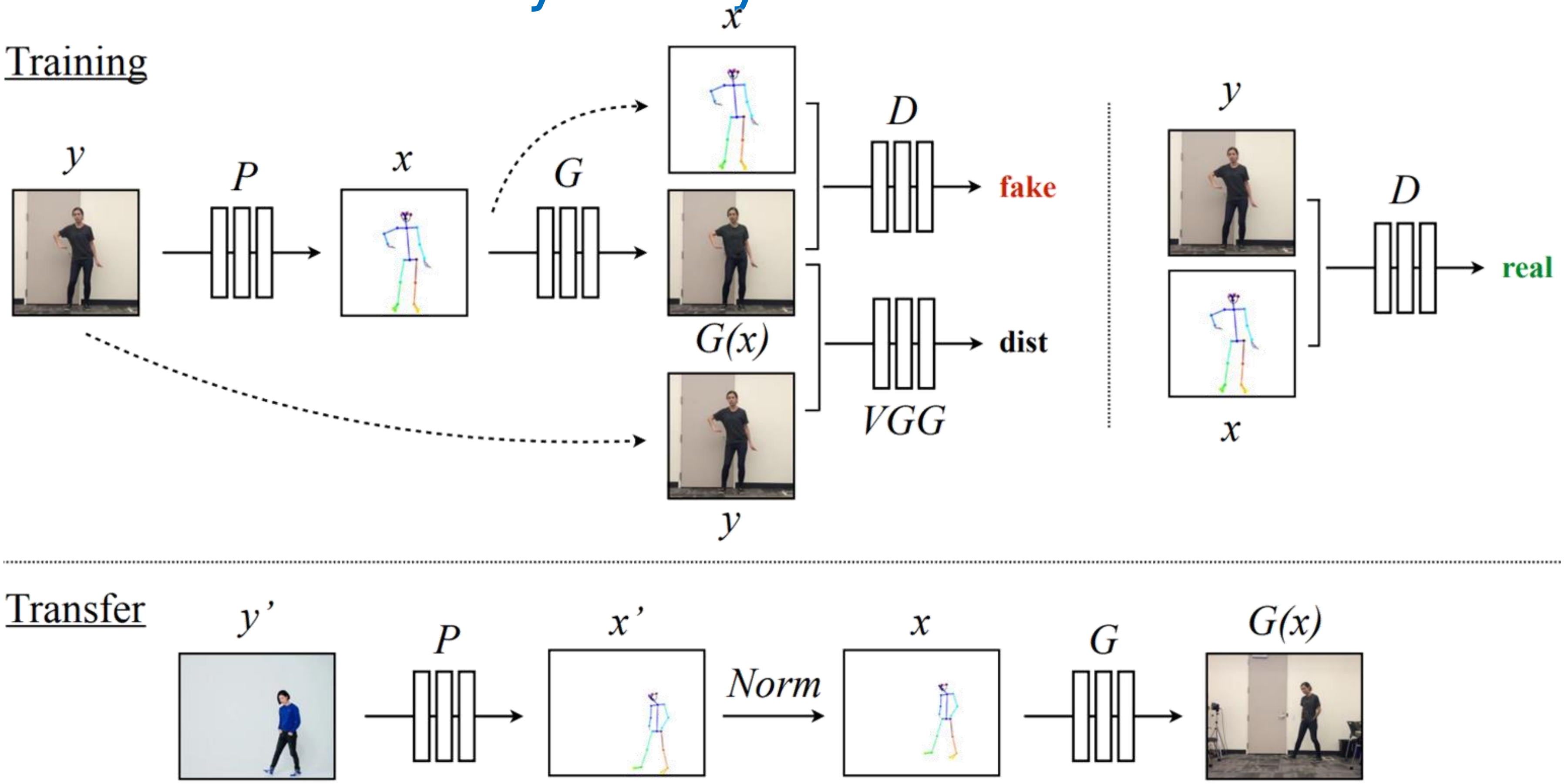
Deep Video Portraits: Insights

- Synthetic data for tracking is great anchor / stabilizer
- Overfitting on small datasets works pretty well
- Need to stay within training set w.r.t. motions
- No real learning; essentially, optimizing the problem with SGD
  -> should be pretty interesting for future directions

Everybody Dance Now

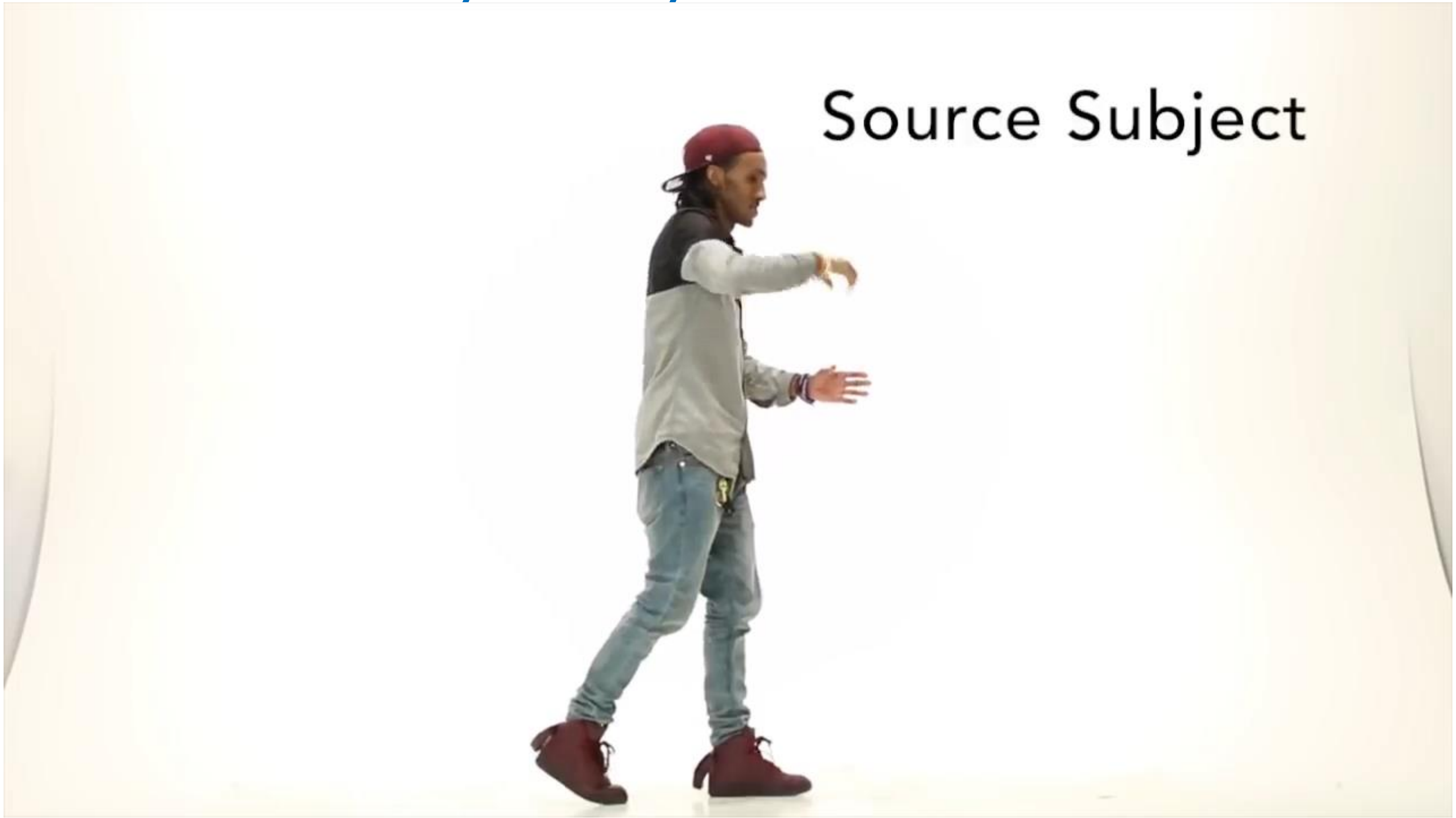
[Chan et al. '18] Everybody Dance Now

Everybody Dance Now



[Chan et al. '18] Everybody Dance Now

Everybody Dance Now



[Chan et al. '18] Everybody Dance Now

Everybody Dance Now

- 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

Everybody Dance Now: Insights

- Conditioning via tracking seems promising!
  - Tracking quality translates to resulting image quality
  - Tracking human skeletons is less developed than faces
    - Temporally it's not stable… (e.g., OpenPose etc.)
  - Fun fact, there were like 4 papers with a similar same idea that appeared around the same time…

Videos still challenging for cGANs…

Pix2Pix [Isola et al. 2017]

Deep Voxels

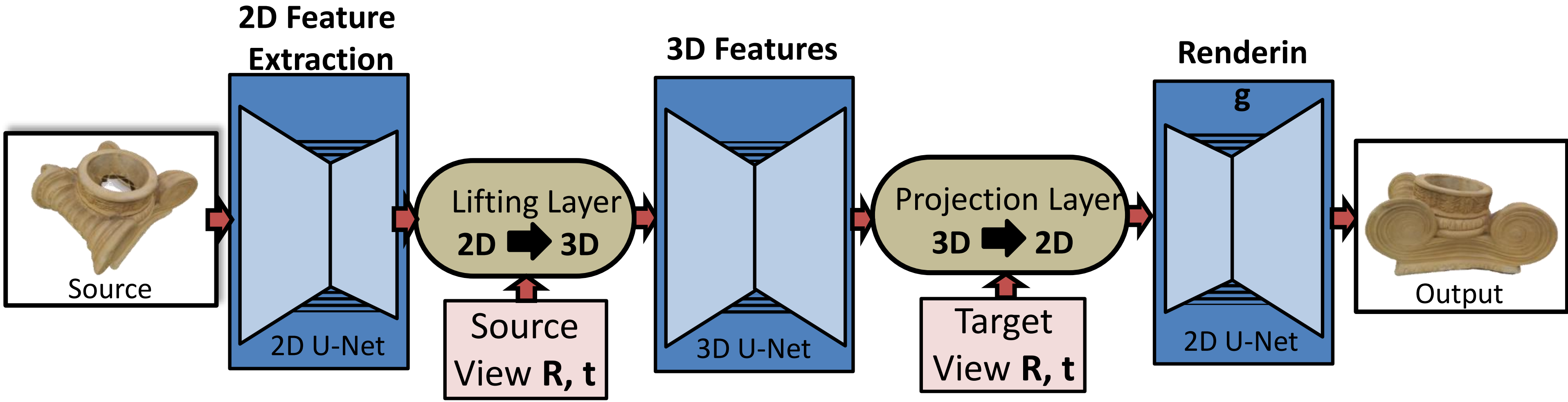
[Sitzmann et al. CVPR'19] Deep Voxels

Deep Voxels

- Main idea for video generation:
  - Why learn 3D operations with 2D Convs !?!?
  - We know how 3D transformations work
    - E.g., 6 DoF rigid pose [R | t]
  - Incorporate these into the architectures
    - Need to be differentiable!
  - Example application: novel view point synthesis
    - Given rigid pose, generate image for that view

[Sitzmann et al. CVPR'19] Deep Voxels

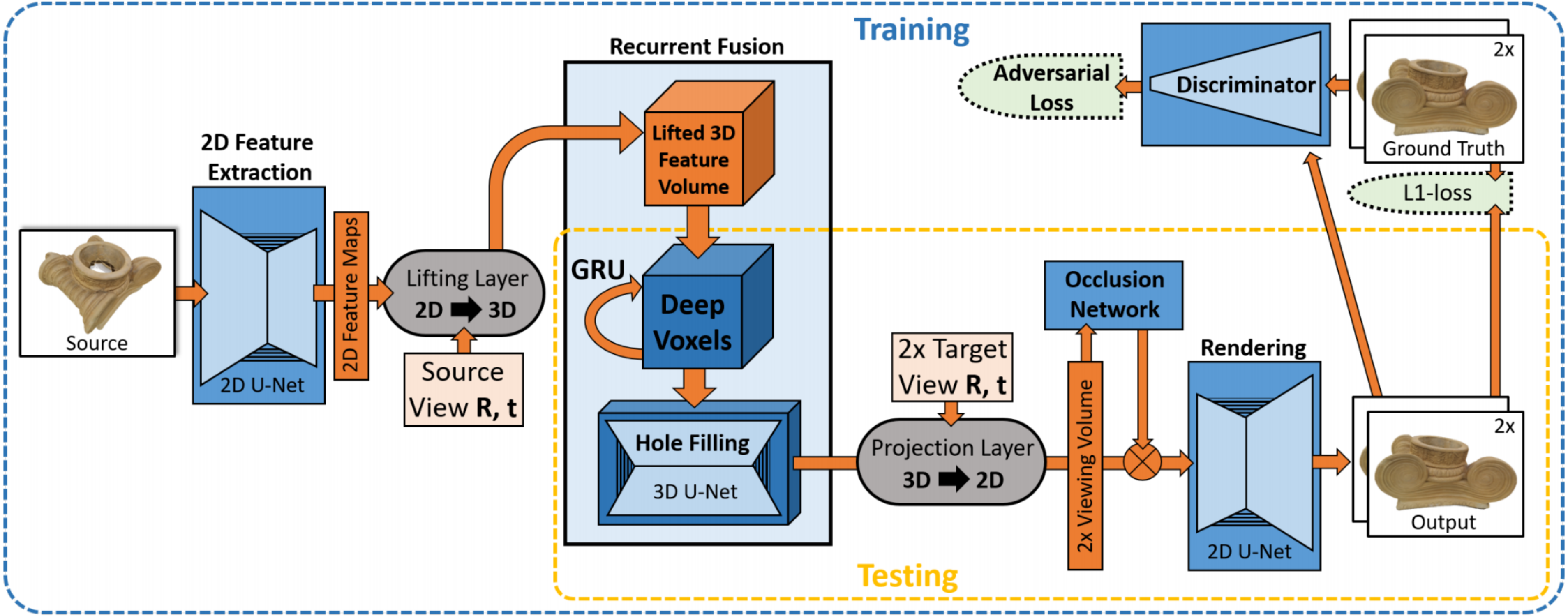
Deep Voxels



Simplified overview for novel view synthesis

[Sitzmann et al. CVPR'19] Deep Voxels

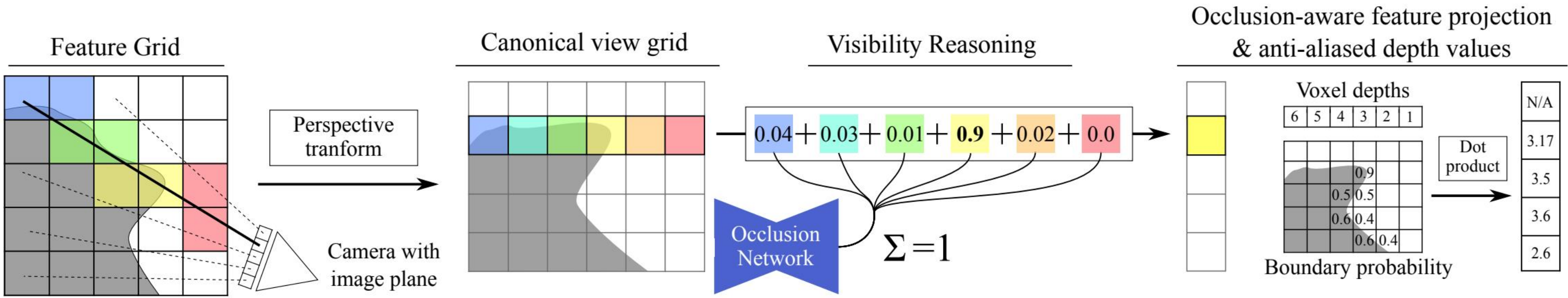
Deep Voxels



[Sitzmann et al. CVPR'19] Deep Voxels

Deep Voxels

Occlusion Network:



Issue: we don't know the depth for the target!

-> Per-pixel softmax along the ray

-> Network learns the depth

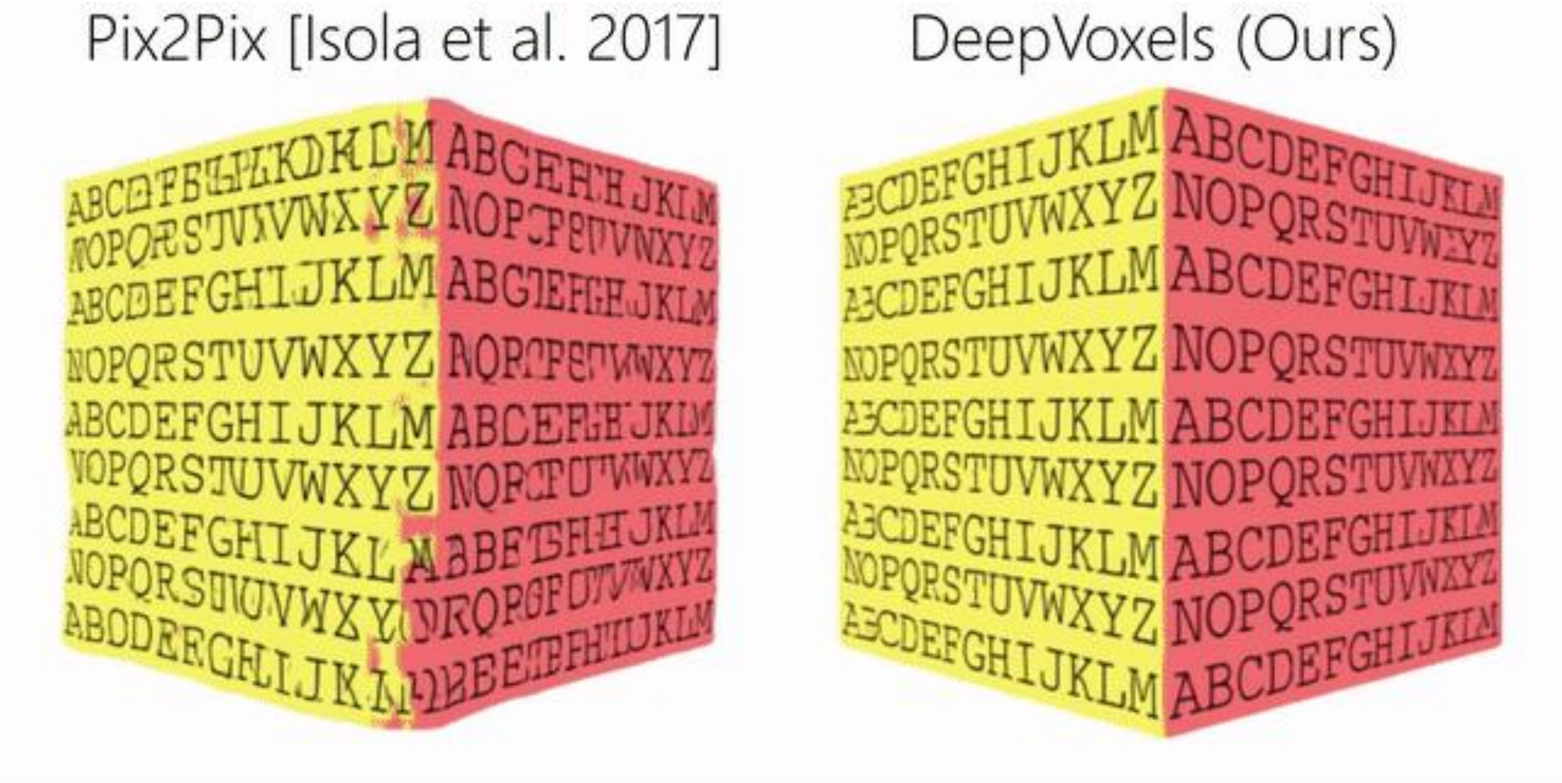
[Sitzmann et al. CVPR'19] Deep Voxels

Deep Voxels

DeepVoxels

Best Baseline: Pix2Pix [Isola et al. 2017]

Deep Voxels



[Sitzmann et al. '18] Deep Voxels

Deep Voxels: Insights

- Lifting from 2D to 3D works great
  - No need to take specific care for temp. coherency!
- All 3D operations are differentiable
- Currently, only for novel view-point synthesis
  - I.e., cGAN for new pose in a given scene
- But: limited resolution due to dense 3D voxel grid

[Sitzmann et al. '18] Deep Voxels

Neural Textures: Features on 3D Mesh

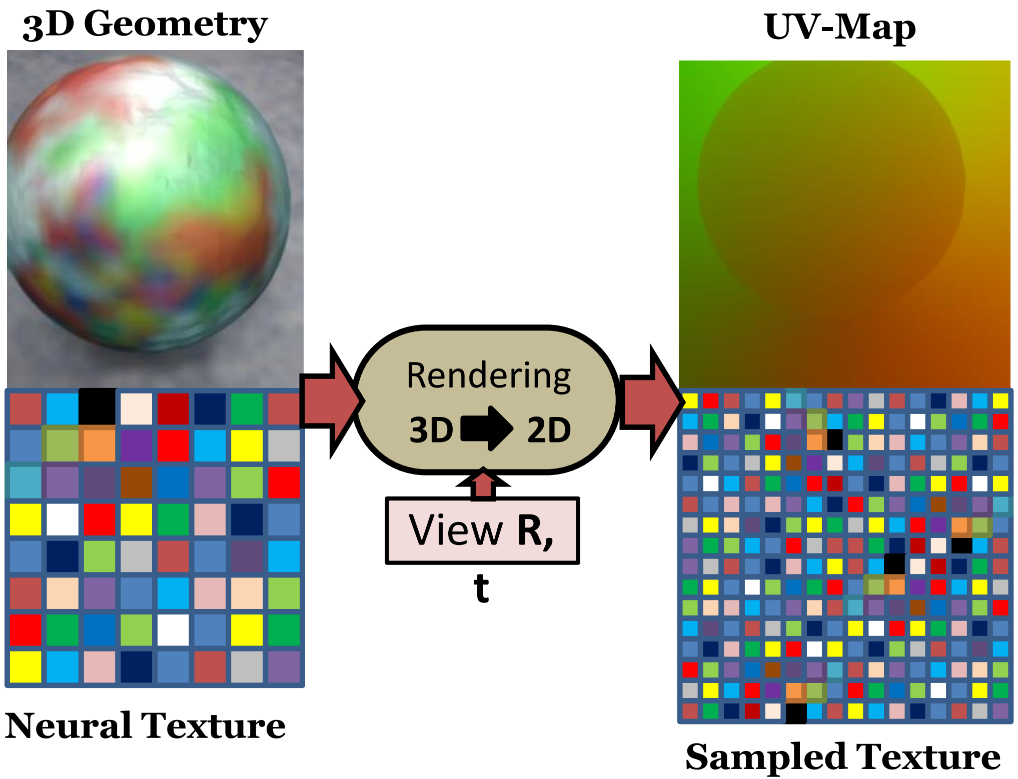
Neural Textures: Features on 3D Mesh

3D Geometry

Neural Texture

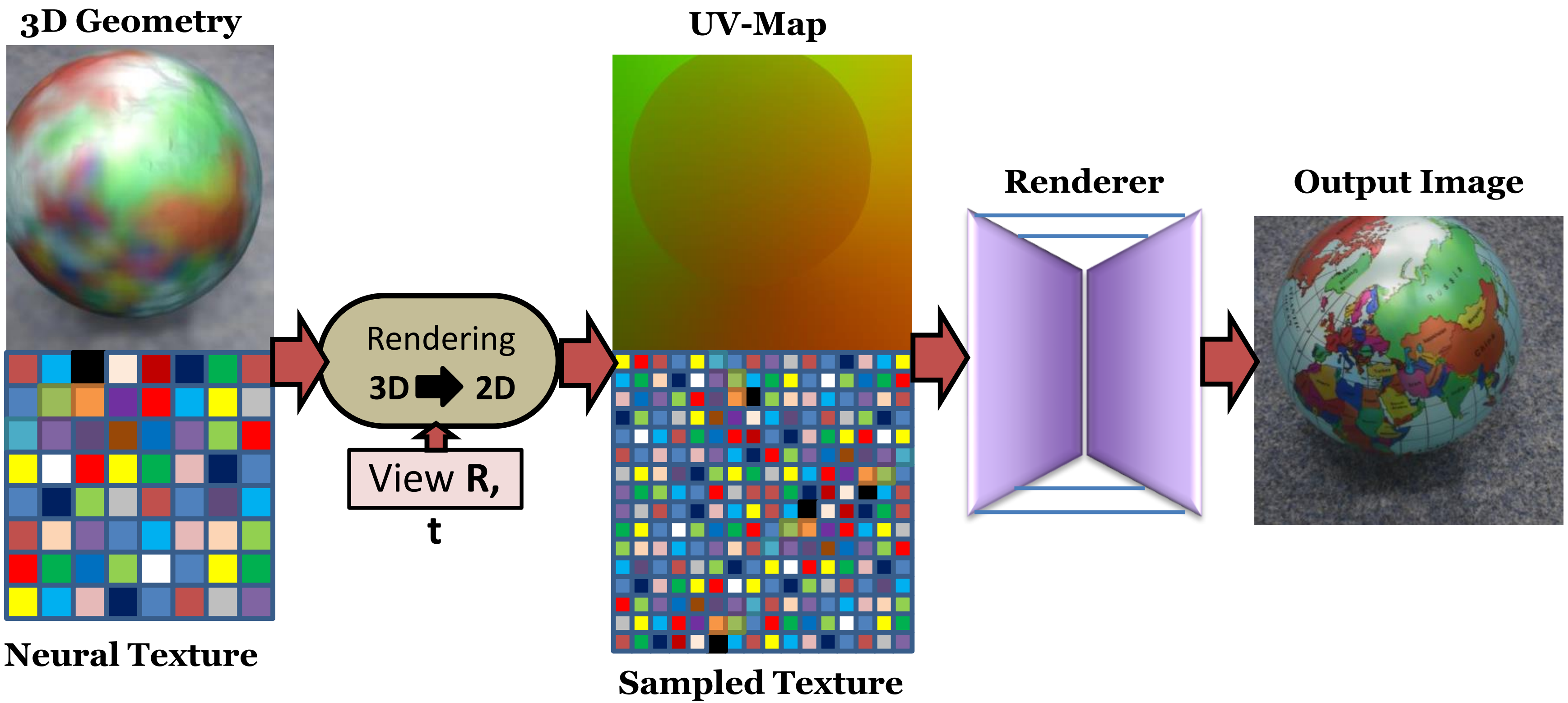
Siggraph'19 [Thies et al.]: Neural Textures

Neural Textures: Features on 3D Mesh



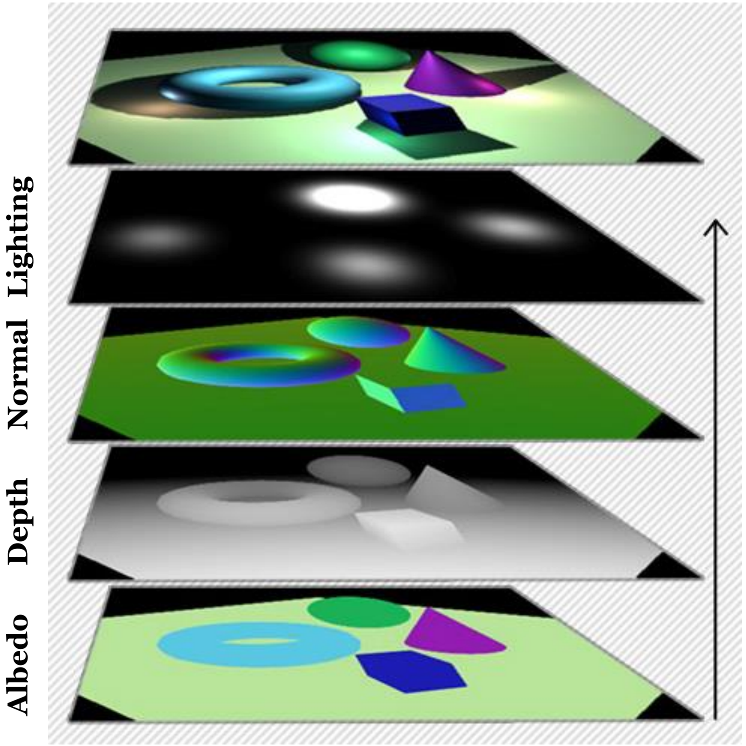
Siggraph'19 [Thies et al.]: Neural Textures

Neural Textures: Features on 3D Mesh



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

Deferred Neural Rendering

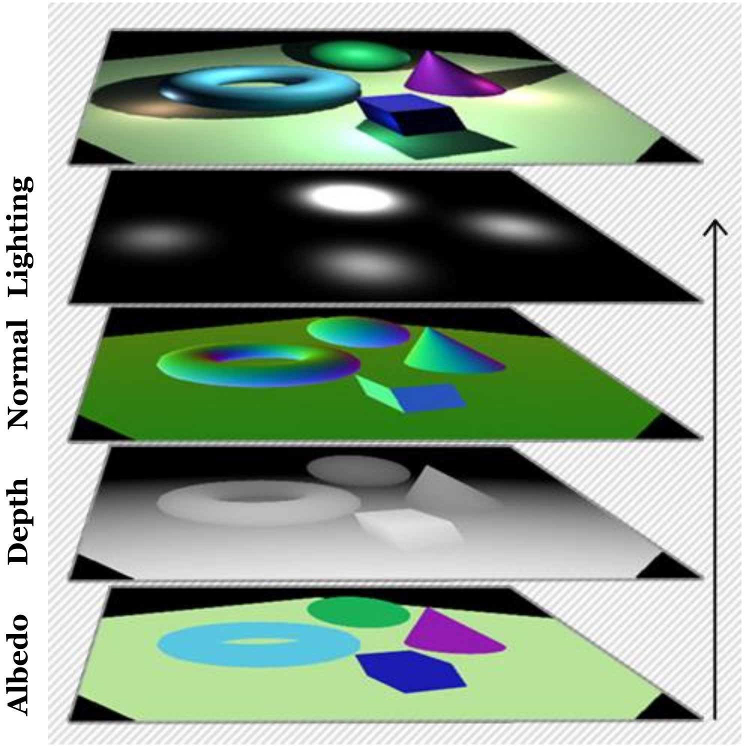


Deferred Renderer

Handcrafted "Feature Maps"

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

Deferred Neural Rendering

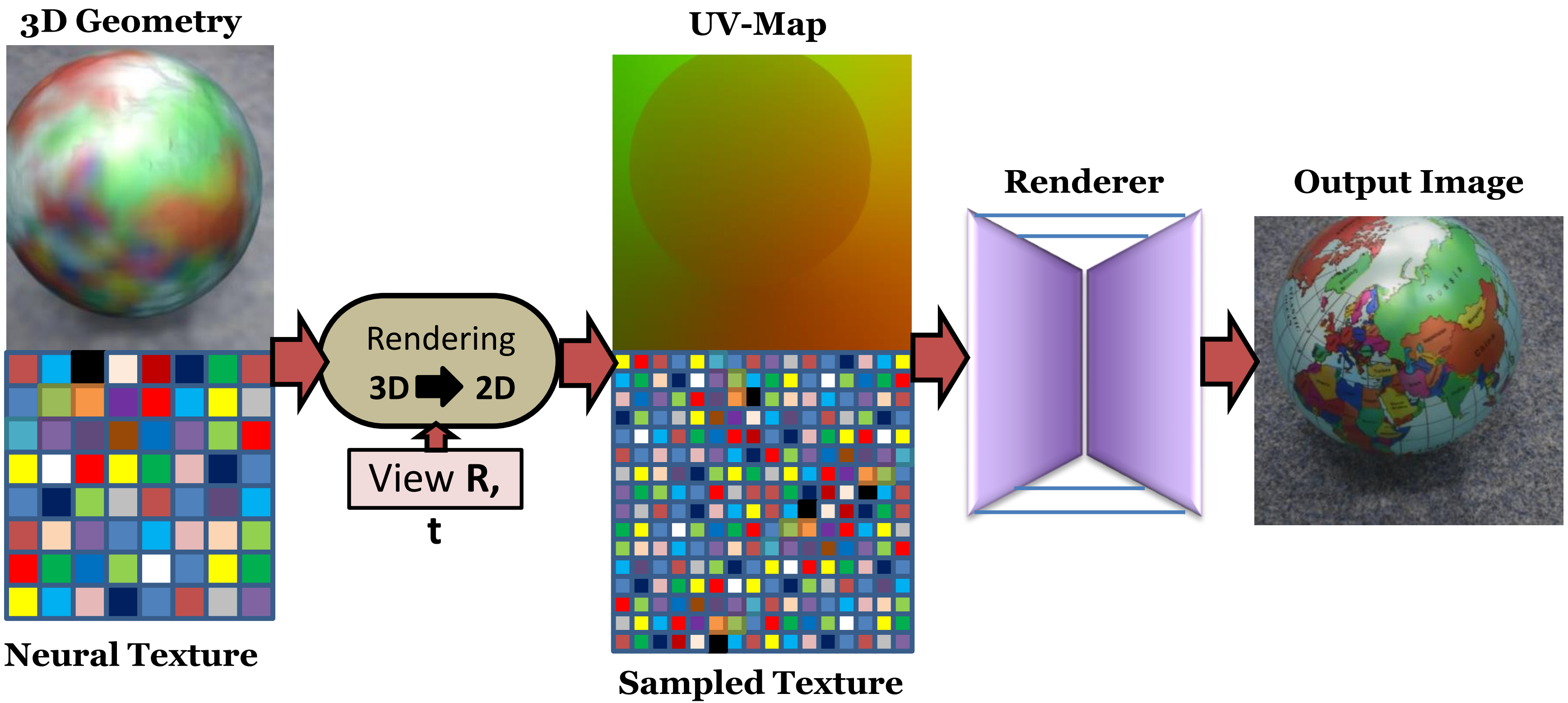


Neural
Deferred Renderer

Learned
~~Handcrafted~~ ”Feature Maps“

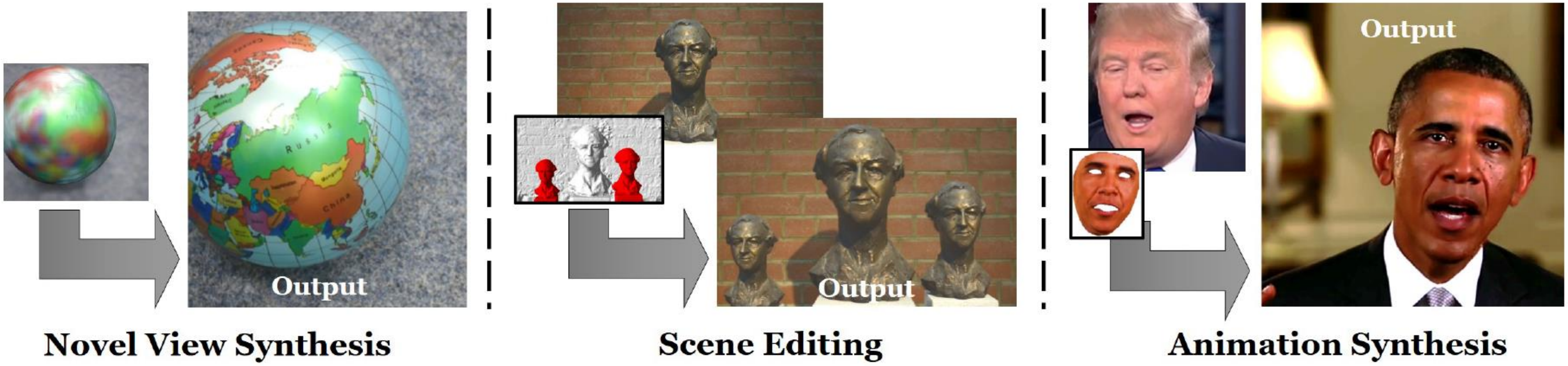
Siggraph’19 [Thies et al.]: Neural Textures

Deferred Neural Rendering



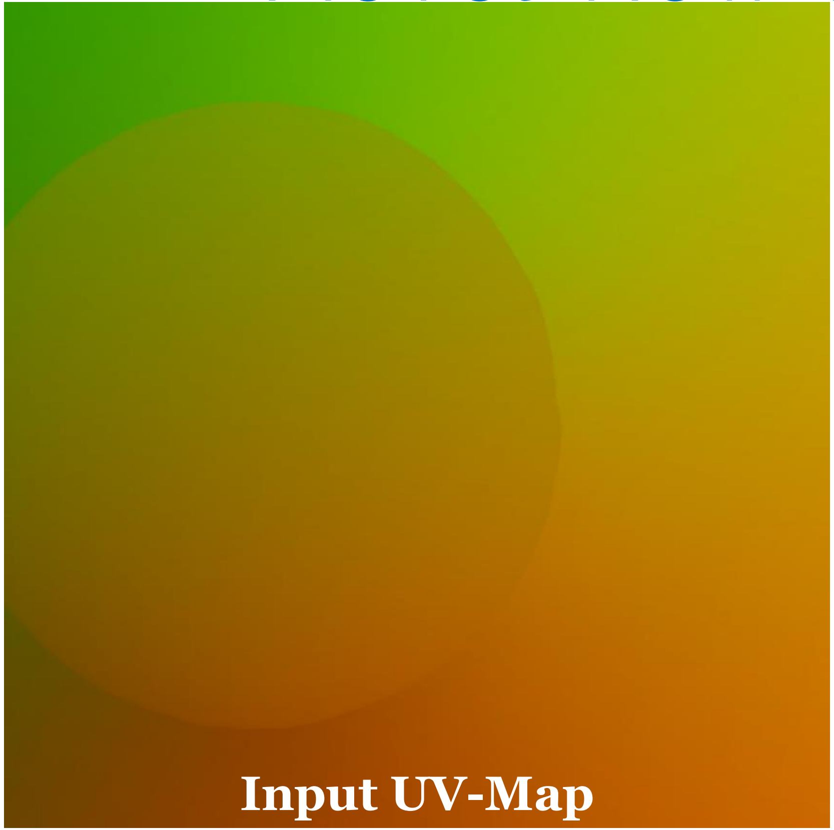
Siggraph'19 [Thies et al.]: Neural Textures

Neural Textures: Features on 3D Mesh



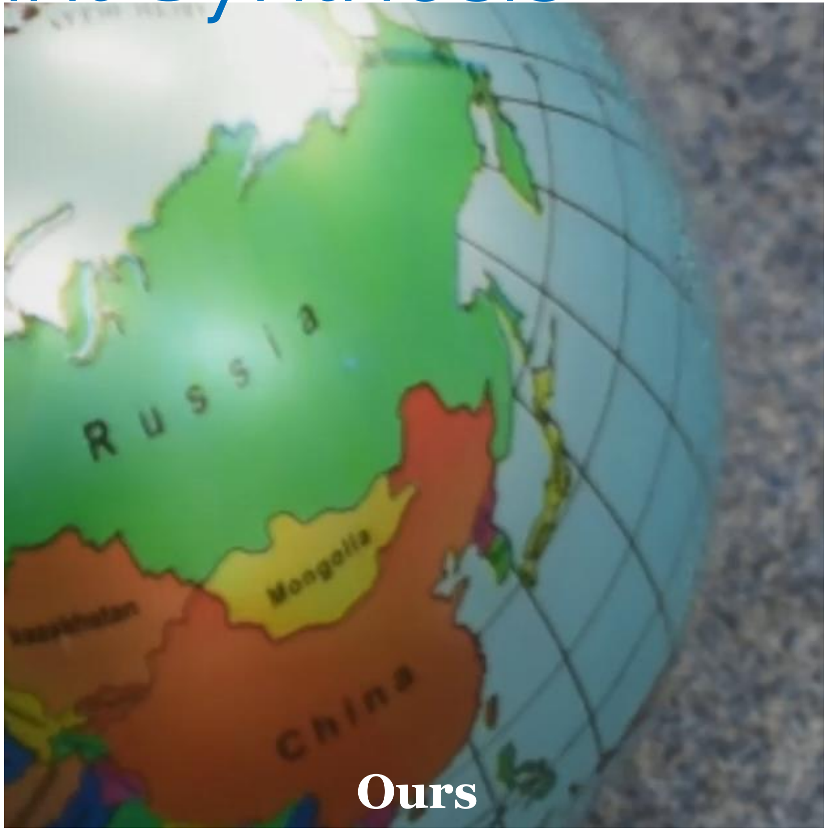
Siggraph’19 [Thies et al.]: Neural Textures

Novel View-Point Synthesis



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

Novel View-Point Synthesis



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

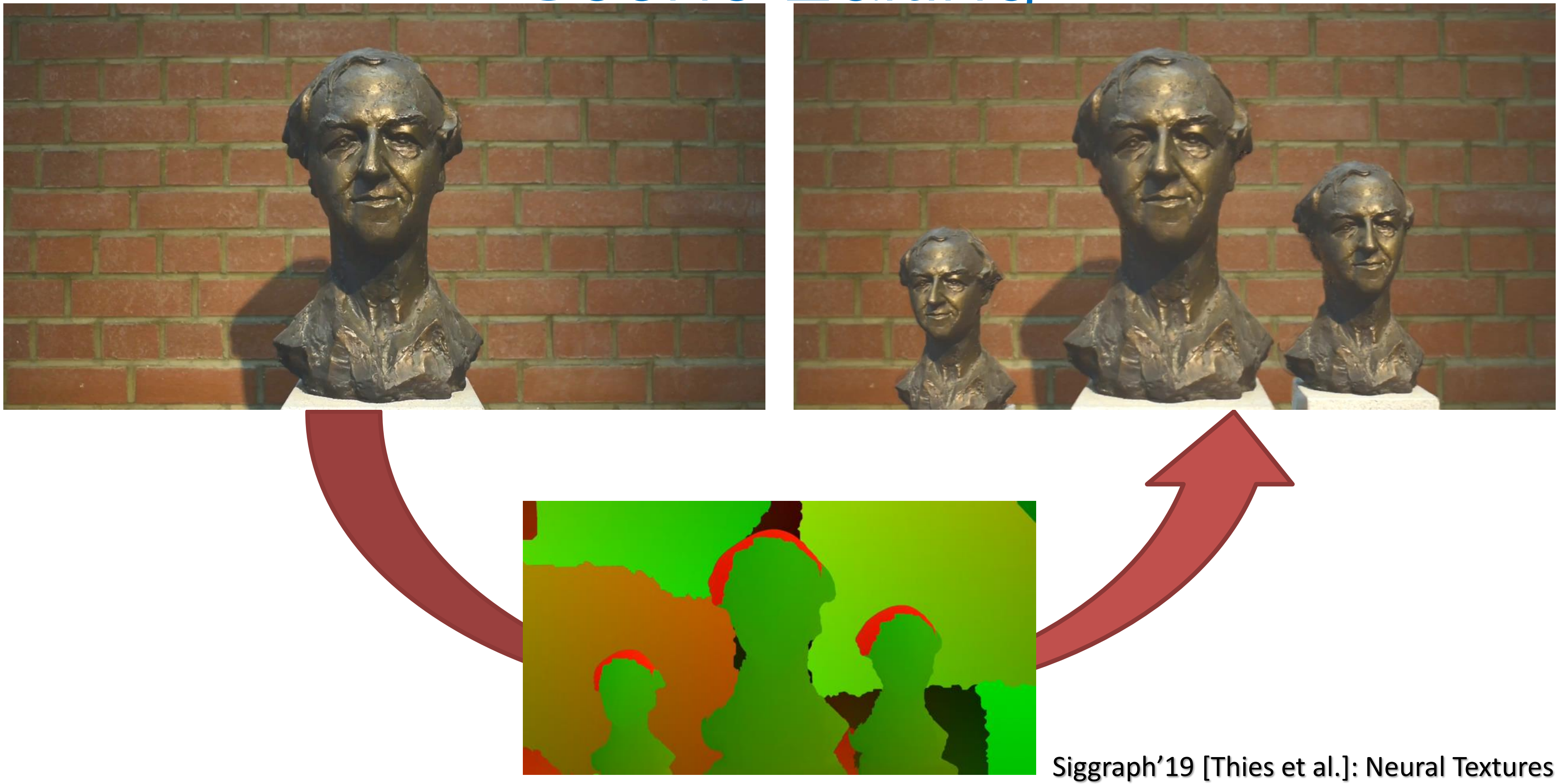
Scene Editing

Input Sequence

Geometry Editing

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

Scene Editing

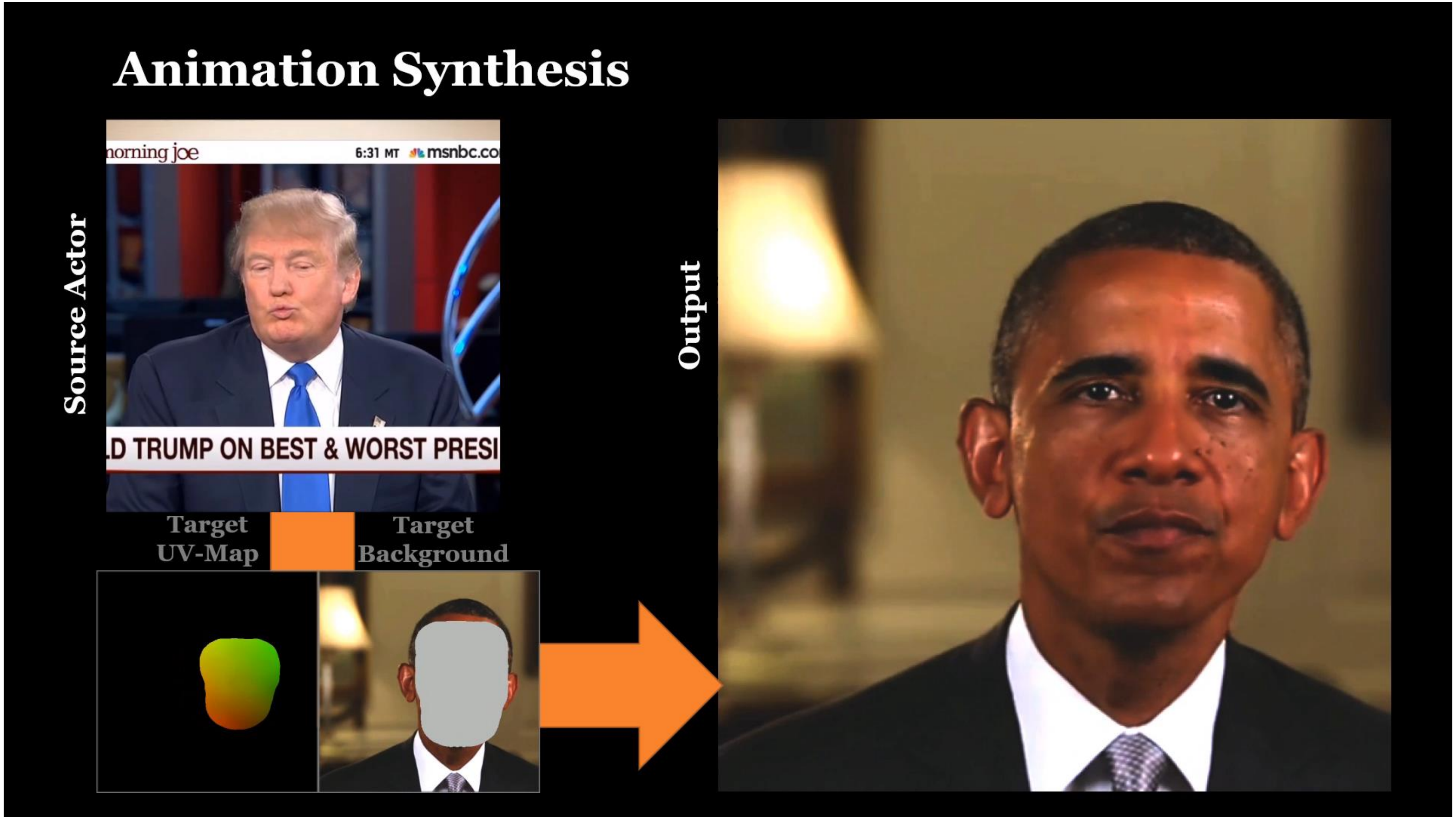


Siggraph'19 [Thies et al.]: Neural Textures

Scene Editing

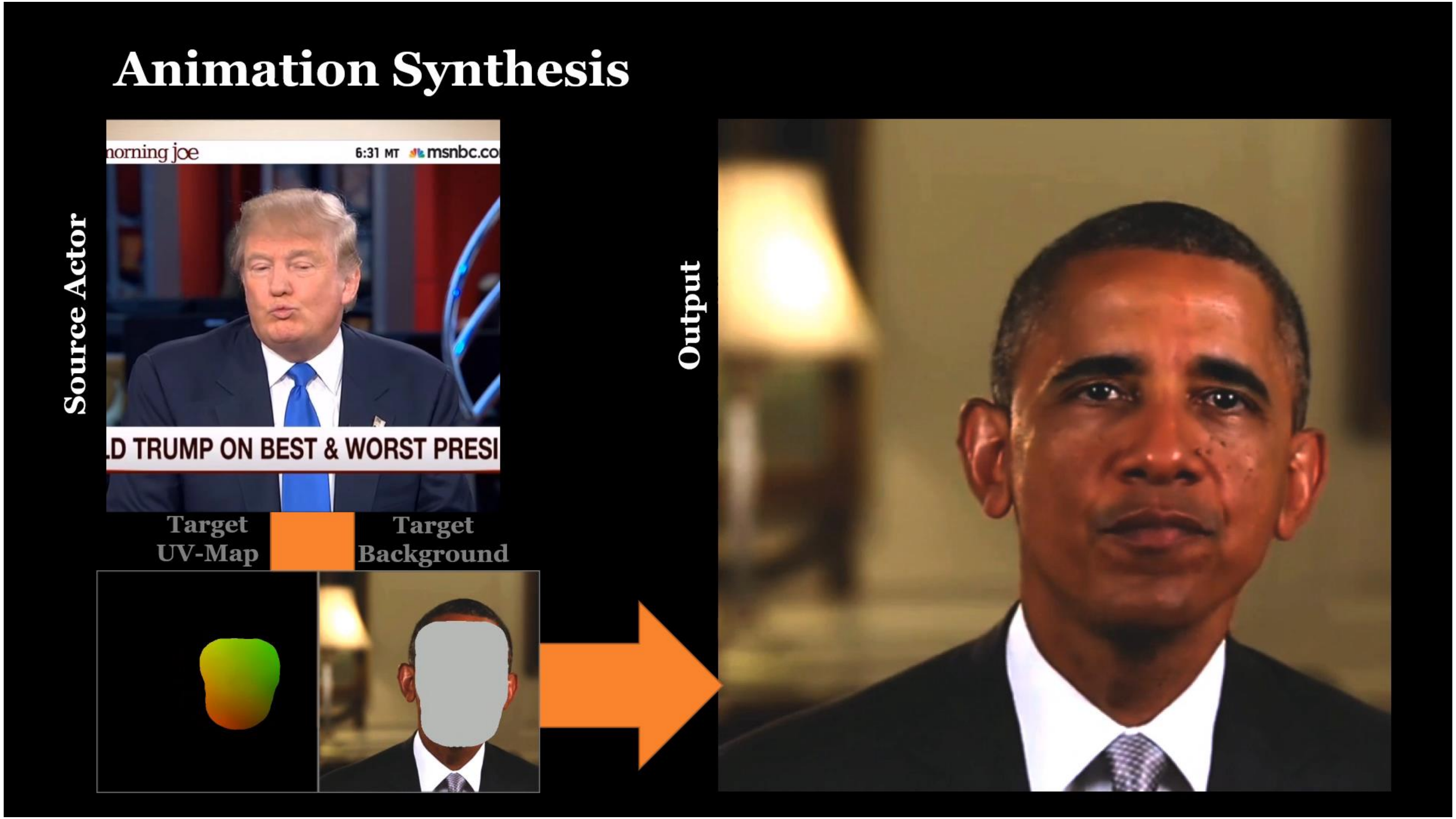
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



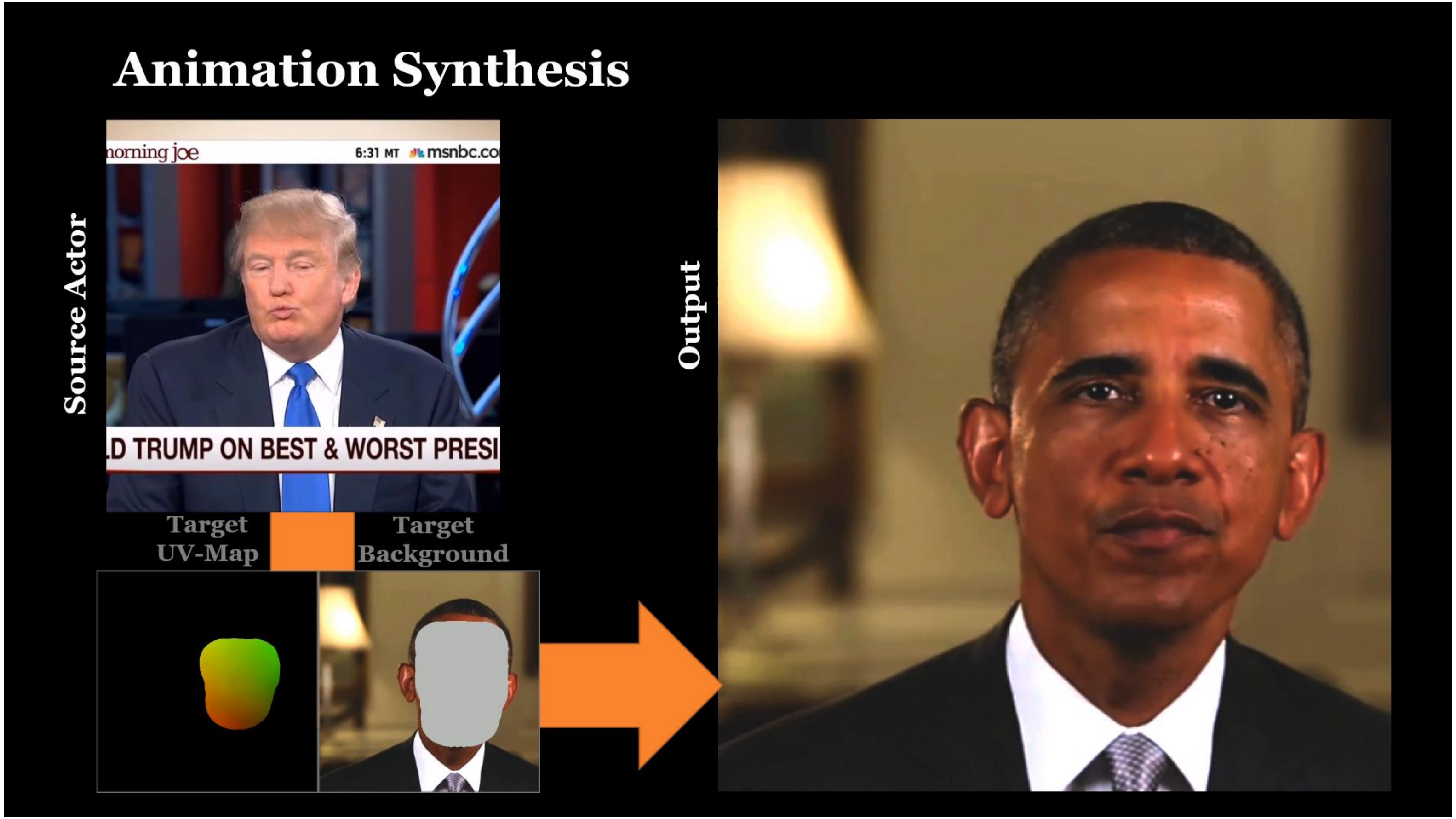
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



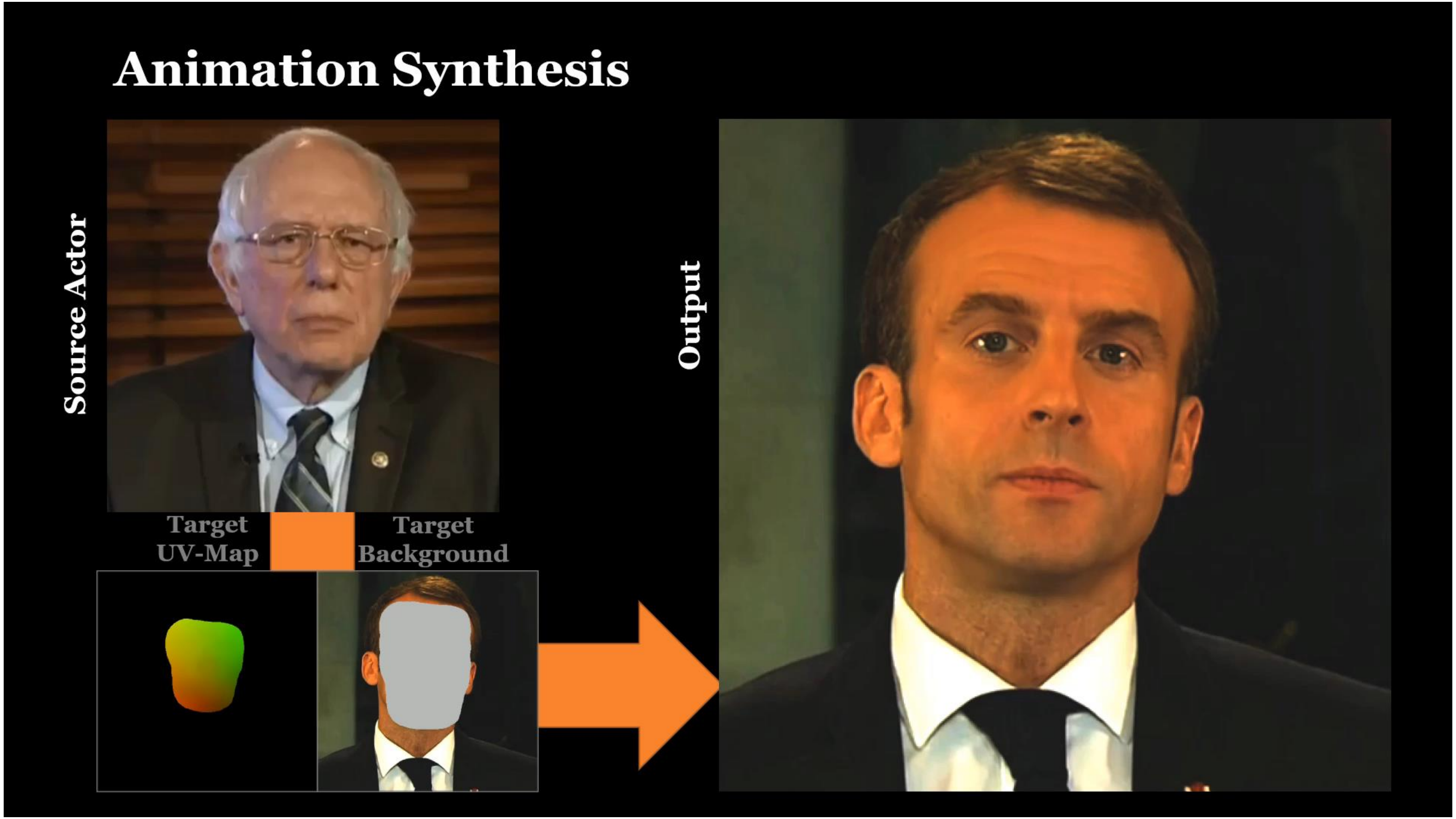
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



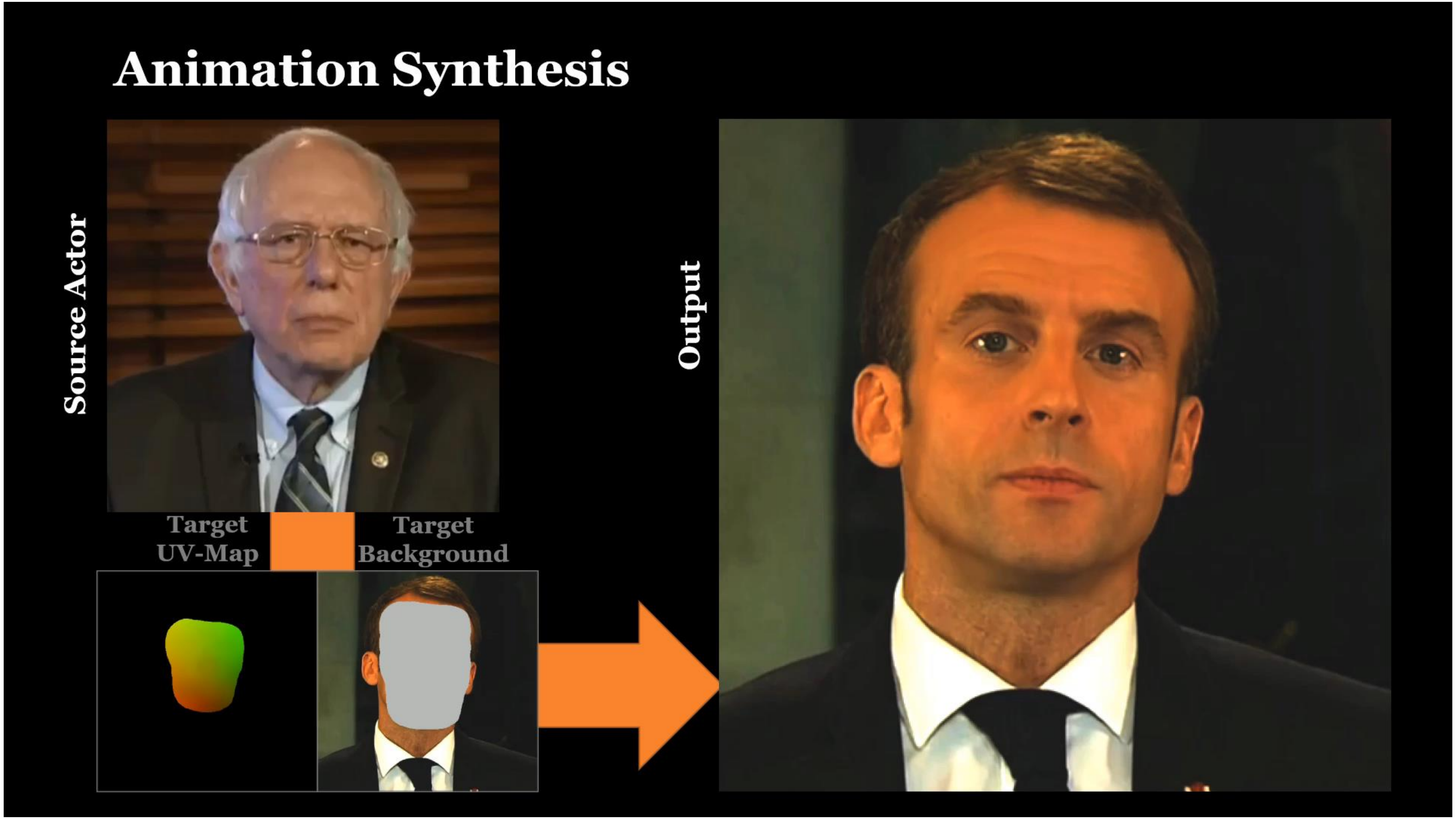
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



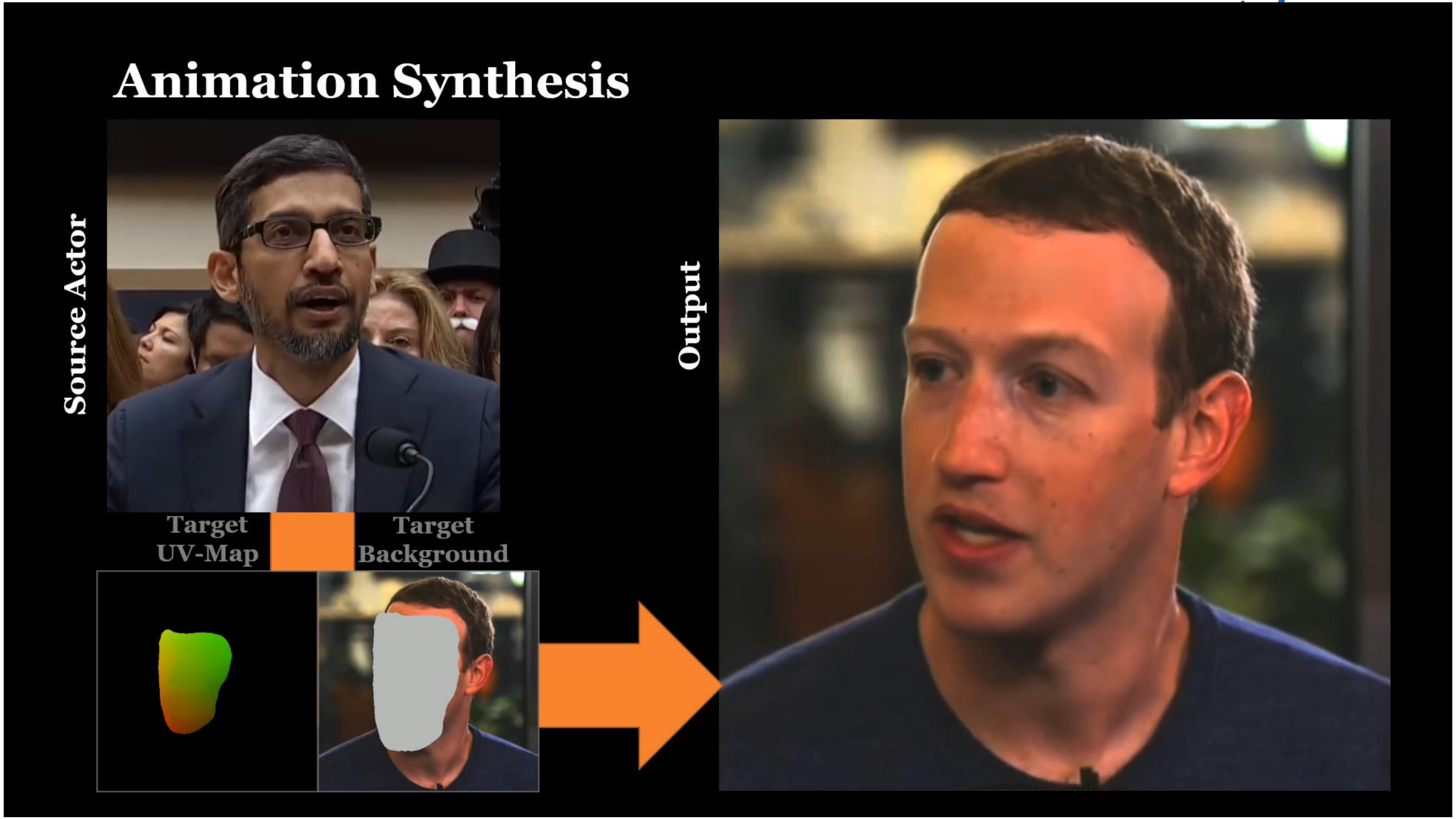
Siggraph'19 [Thies et al.]: Neural Textures

Facial Animation



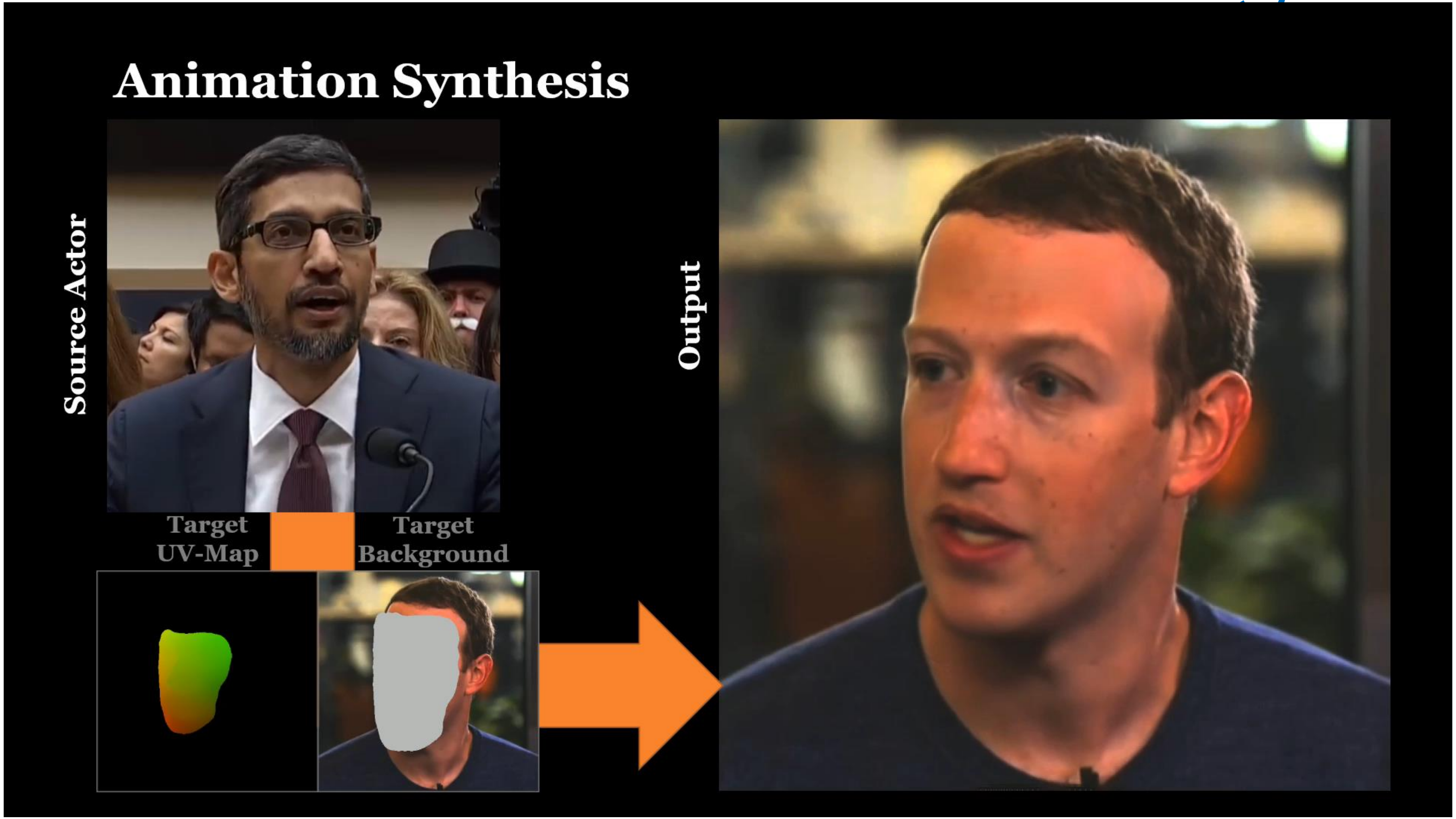
Siggraph'19 [Thies et al.]: Neural Textures

Deferred Neural Rendering



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

Deferred Neural Rendering



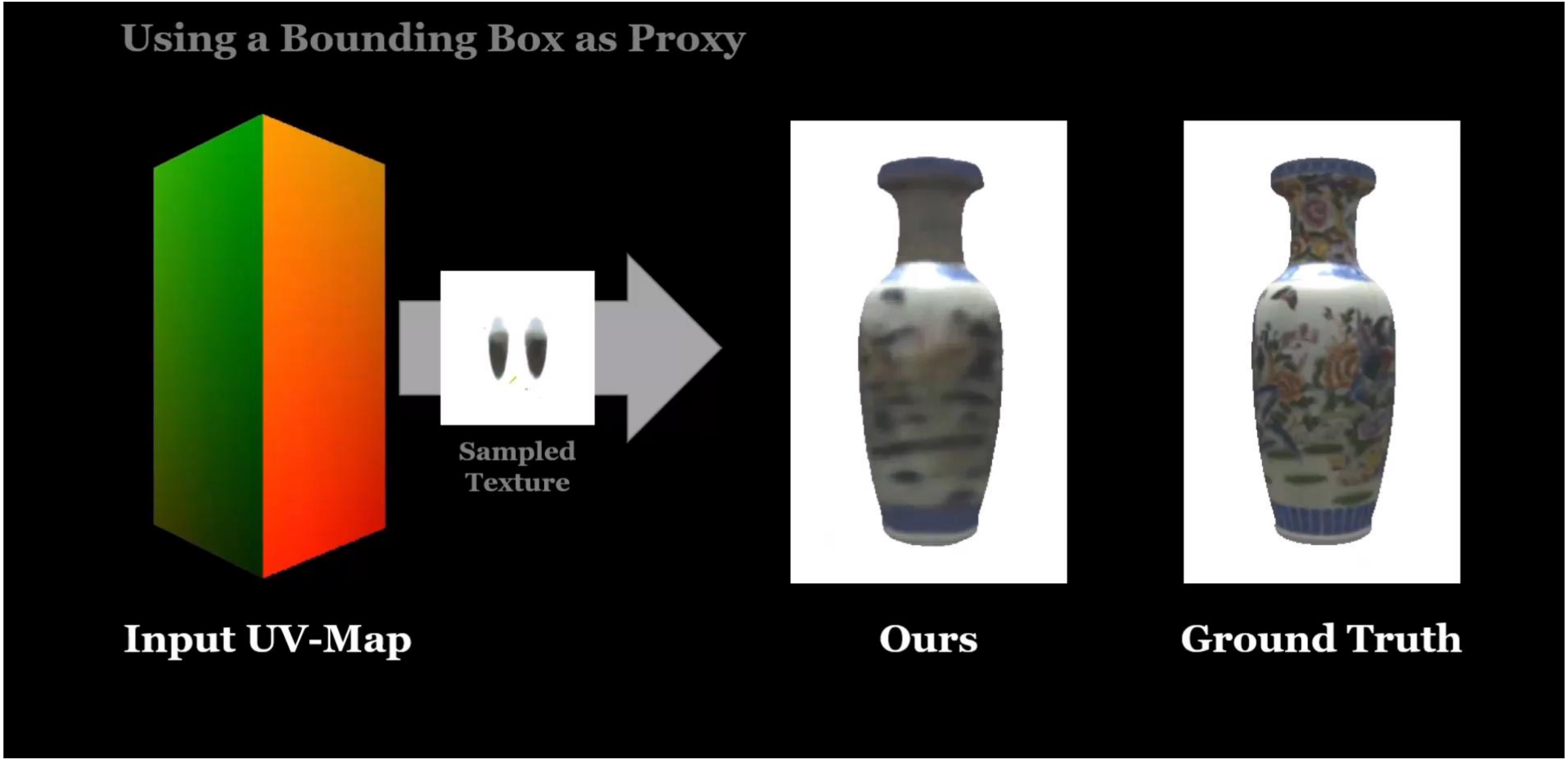
Siggraph'19 [Thies et al.]: Neural Textures

Big Open Challenges

Big Open Challenges

Photo-realistic Reconstruction

Big Open Challenges: How much can AI do?

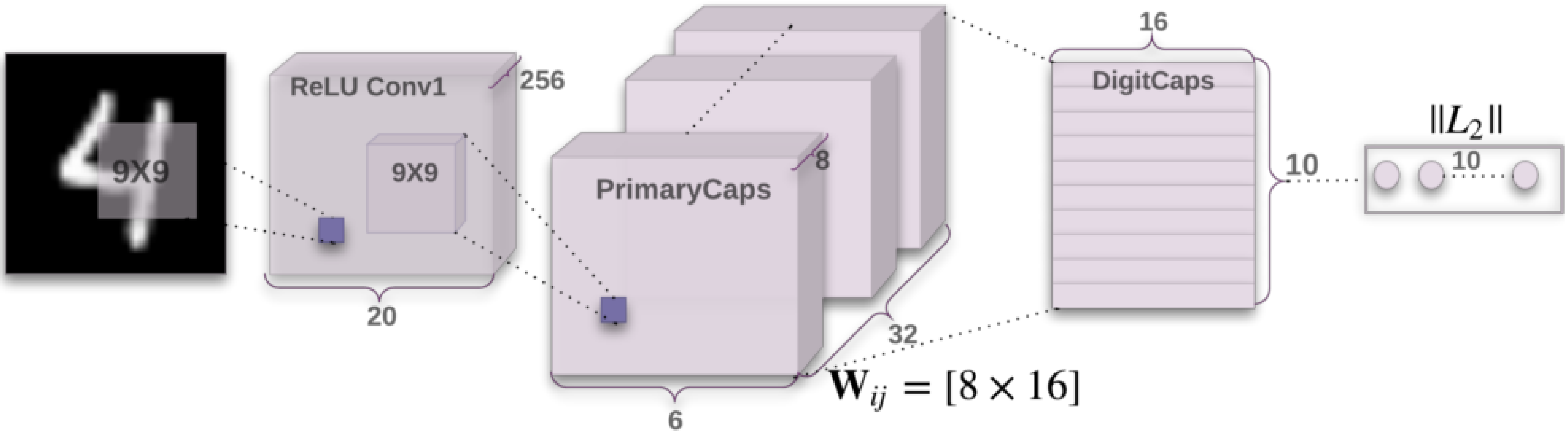


Siggraph'19 [Thies et al.]: Neural Textures

Big Open Challenges: 3D in Networks

Why learn 3D operations, such as transformations?

-> *differentiate known operators*



Capsule networks are motivated by *inverse graphics* [Sabour et al. 17]

Autoregressive Models

Autoregressive Models vs GANs

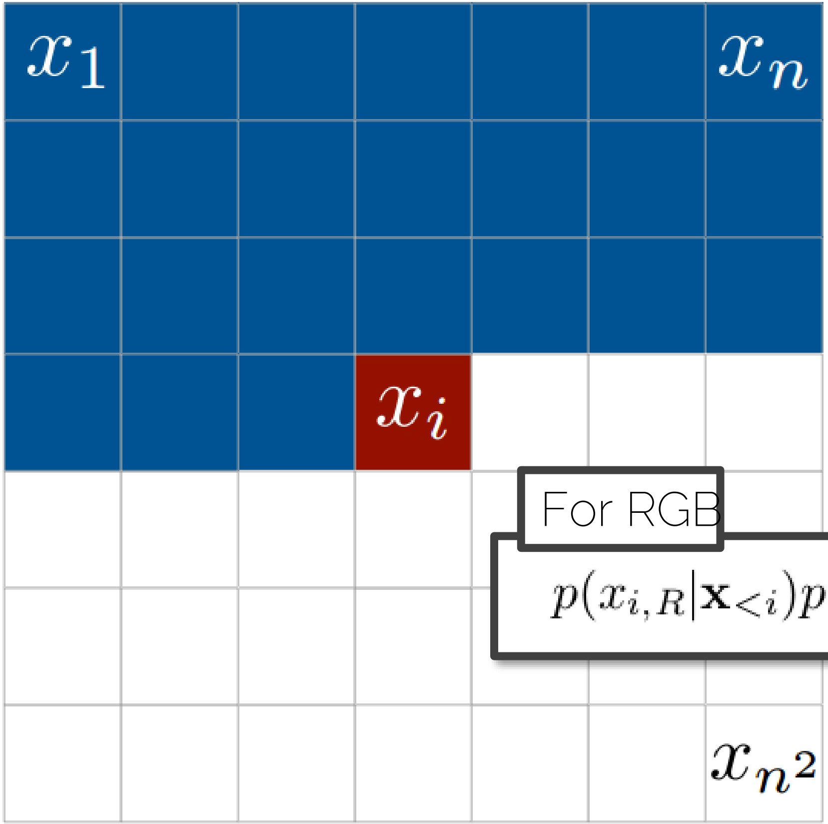
- GANs learn implicit data distribution
  - i.e., output are samples (distribution is in model)
- Autoregressive models learn an explicit distribution governed by a prior imposed by model structure
  - i.e., outputs are probabilities (e.g., softmax)

PixelRNN

- Goal: model distribution of natural images
- Interpret pixels of an image as product of conditional distributions
  - Modeling an image → sequence problem
  - Predict one pixel at a time
  - Next pixel determined by all previously predicted pixels
  - ➢ Use a Recurrent Neural Network

[Van den Oord et al 2016]

PixelRNN



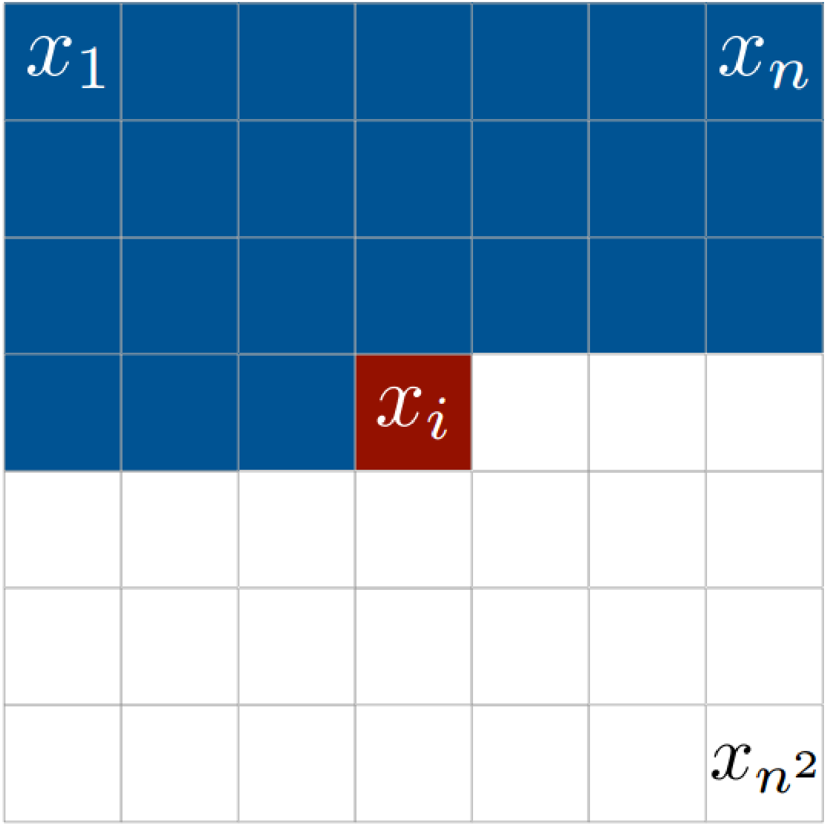
$$p(\mathbf{x}) = \prod_{i=1}^{n^2} p(x_i | x_1, ..., x_{i-1})$$

For RGB

$$p(x_{i,R}|\mathbf{x}_{<i})p(x_{i,G}|\mathbf{x}_{<i}, x_{i,R})p(x_{i,B}|\mathbf{x}_{<i}, x_{i,R}, x_{i,G})$$

[Van den Oord et al 2016]

PixelRNN



$$p(\mathbf{x}) = \prod_{i=1}^{n^2} p(x_i | x_1, \ldots, x_{i-1})$$

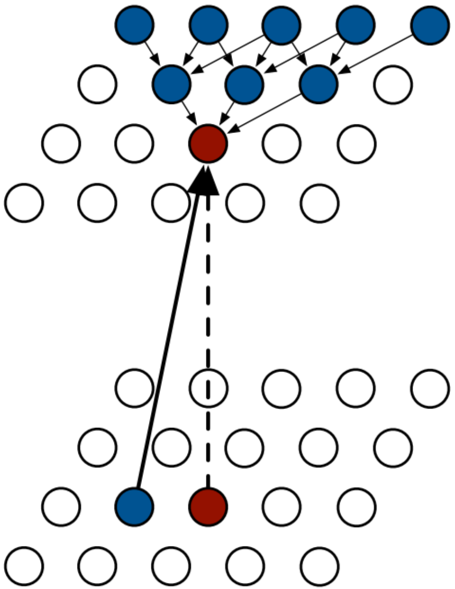
$x_i \in [0,255]$

→ 256-way softmax

[Van den Oord et al 2016]

PixelRNN

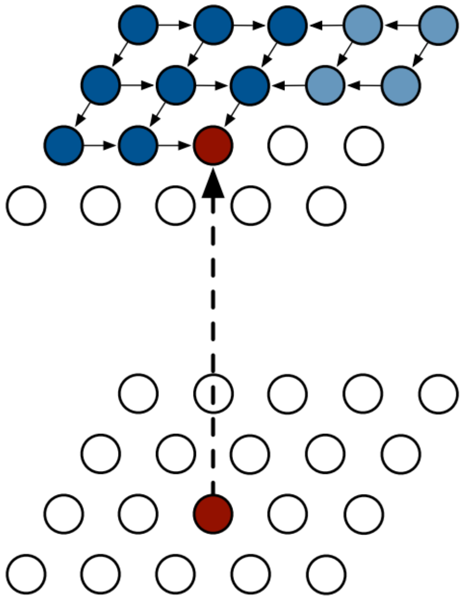
- Row LSTM model architecture
- Image processed row by row
- Hidden state of pixel depends on the 3 pixels above it
  - Can compute pixels in row in parallel
- Incomplete context for each pixel



[Van den Oord et al 2016]

PixelRNN

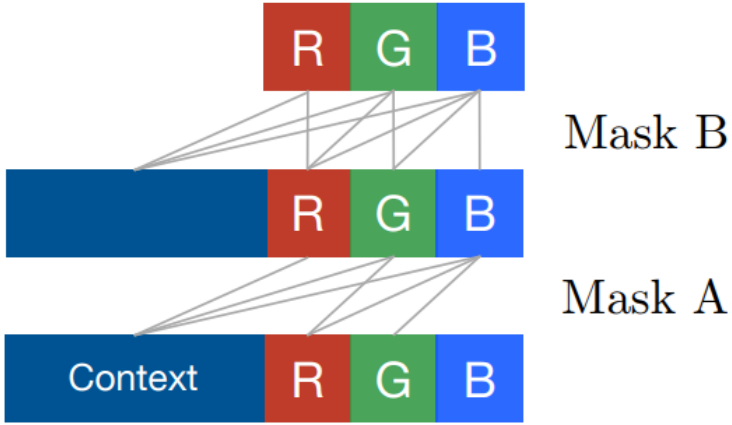
- Diagonal BiLSTM model architecture
- Solve incomplete context problem
- Hidden state of pixel $p_{i,j}$ depends on $p_{i,j-1}$ and $p_{i-1,j}$
- Image processed by diagonals



[Van den Oord et al 2016]

PixelRNN

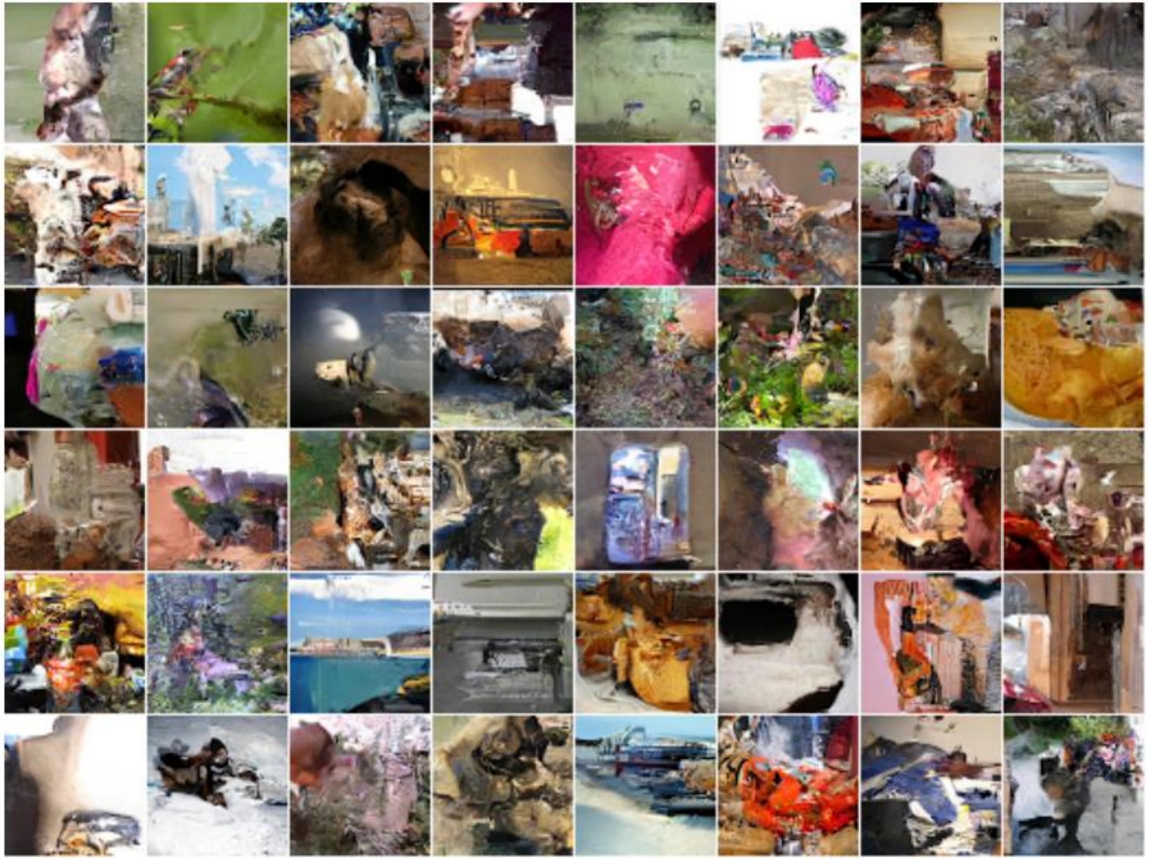
- Masked Convolutions
- Only previously predicted values can be used as context
- Mask A: restrict context during 1st conv
- Mask B: subsequent convs
- Masking by zeroing out values



[Van den Oord et al 2016]

PixelRNN

- Generated 64x64 images, trained on ImageNet



[Van den Oord et al 2016]

PixelCNN

- Row and Diagonal LSTM layers have potentially unbounded dependency range within the receptive field
  - Can be very computationally costly

➢ PixelCNN:
  - standard convs capture a bounded receptive field
  - All pixel features can be computed at once (during training)

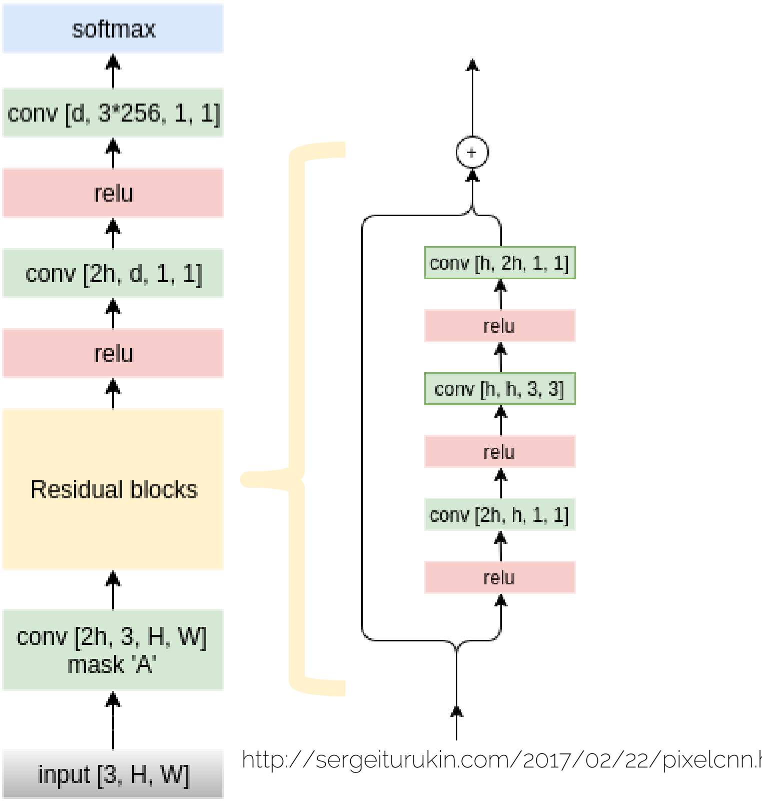
[Van den Oord et al 2016]

PixelCNN

- Model preserves spatial dimensions
- Masked convolutions to avoid seeing future context

1	1	1	1	1
1	1	1	1	1
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0

Mask A



http://sergeiturukin.com/2017/02/22/pixelcnn.

[Van den Oord et al 2016]

Gated PixelCNN

- Gated blocks
- Imitate multiplicative complexity of PixelRNNs to reduce performance gap between PixelCNN and PixelRNN
- Replace ReLU with gated block of sigmoid, tanh

k^th layer   sigmoid

$$y = \tanh\left(W_{k,f} * x\right) \odot \sigma\left(W_{k,g} * x\right)$$

element-wise product   convolution

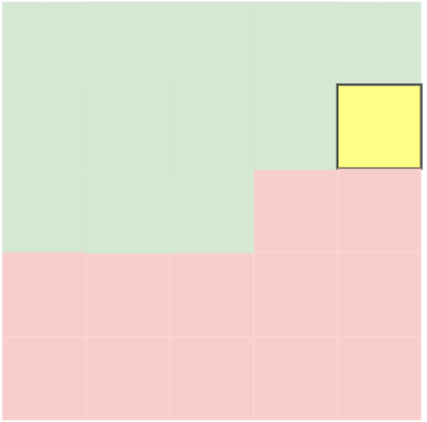
[Van den Oord et al 2016]

PixelCNN Blind Spot

1	1	1	1	1
1	1	1	1	1
1	1	0	0	0
0	0	0	0	0
0	0	0	0	0

5x5 image / 3x3 conv

Receptive Field



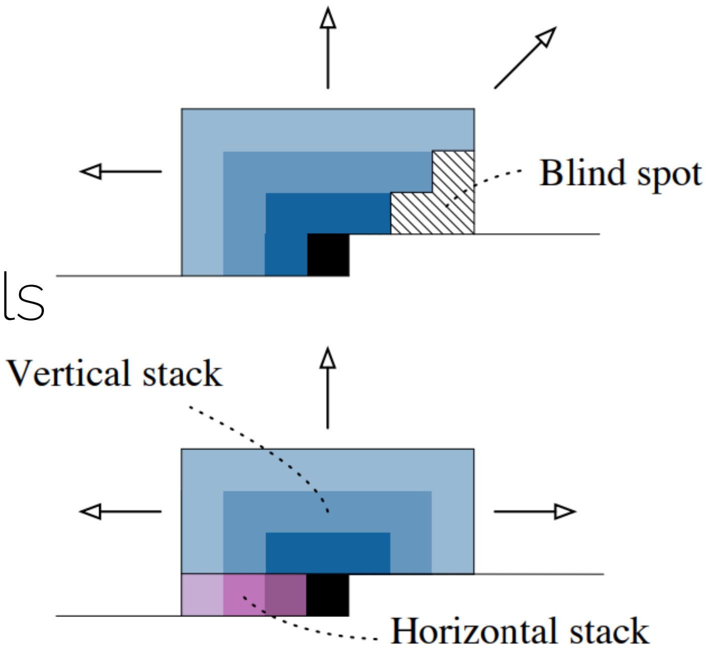
Unseen context

http://sergeiturukin.com/2017/02/24/gated-pixelcnn.html

[Van den Oord et al 2016]

PixelCNN: Eliminating Blind Spot

- Split convolution to two stacks
- Horizontal stack conditions on current row
- Vertical stack conditions on pixels above



[Van den Oord et al 2016]

Conditional PixelCNN

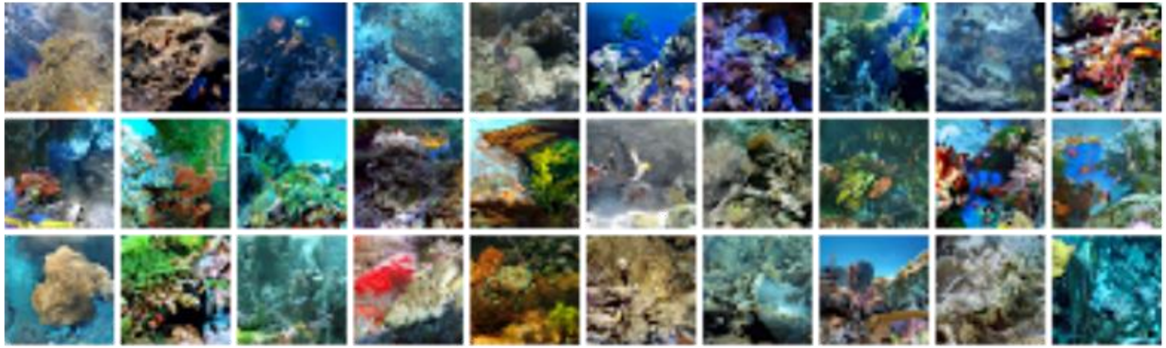
- Conditional image generation
- E.g., condition on semantic class, text description

latent vector to be conditioned on

$$y = \tanh\left(W_{k,f} * x + V_{k,f}^{T} h\right) \odot \sigma(W_{k,g} * x + V_{k,g}^{T} h)$$

[Van den Oord et al 2016]

Conditional PixelCNN



Coral Reef

Sorrel horse

[Van den Oord et al 2016]

Autoregressive Models vs GANs

- Advantages of autoregressive:
  - Explicitly model probability densities
  - More stable training
  - Can be applied to both discrete and continuous data
- Advantages of GANs:
  - Have been empirically demonstrated to produce higher quality images
  - Faster to train

Autoregressive Models

- State of the art is pretty impressive ☺

Vector Quantized Variational AutoEncoder

Generating Diverse High-Fidelity Images with VQ-VAE-2
https://arxiv.org/pdf/1906.00446.pdf [Razavi et al. 19]

See you next week ☺