Training a 3D-Aware GAN

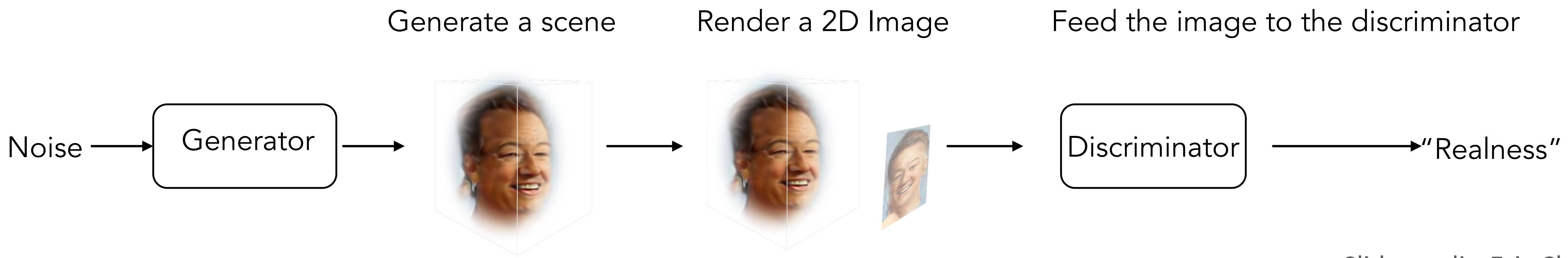
Training Steps

1. Generate a representation of a *scene*

2. Render the scene from a random camera pose

3. Feed the image to a 2D discriminator

4. Backpropagate through the discriminator and differentiable rendering



Slide credit: Eric Chan

pi-GAN: Periodic Implicit Generative Adversarial Networks for 3D-Aware Image Synthesis [Chan et al., 2021]

Text-based Editing

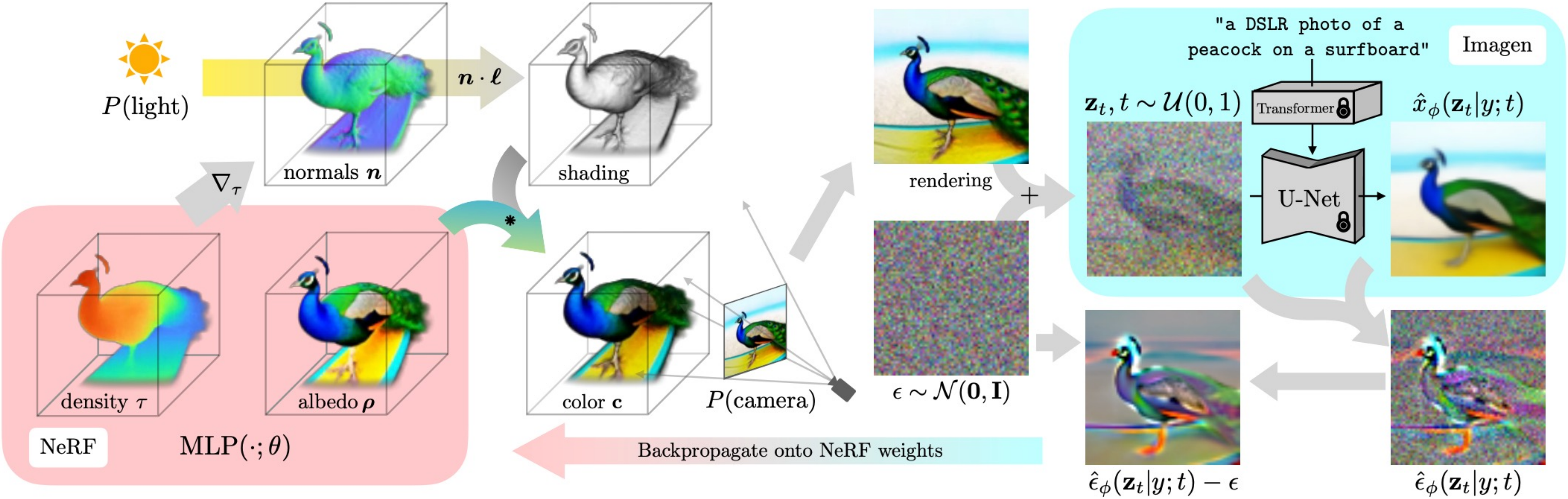
a DSLR photo of a squirrel

wearing a purple hoodie

reading a book

DreamFusion: Text-to-3D using 2D Diffusion [Poole et al., 2022]

Text-based Editing



FOR loop

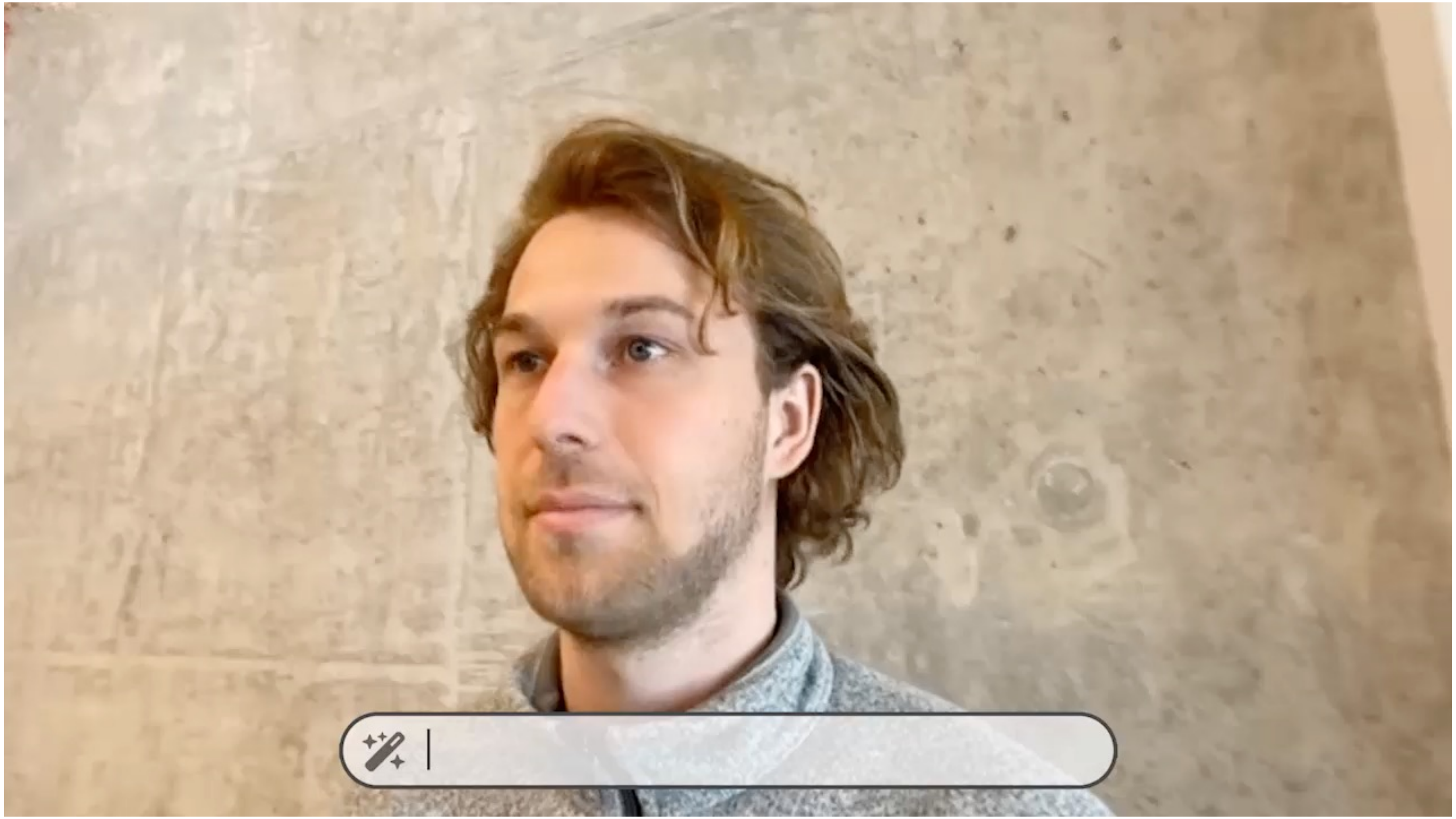
Step 1. Render a view using existing NeRF

Step 2. Add noise and denoise using a pre-trained Stable Diffusion model

Step 3. Update NeRF parameters with the gradient (difference between added and predicted noises)

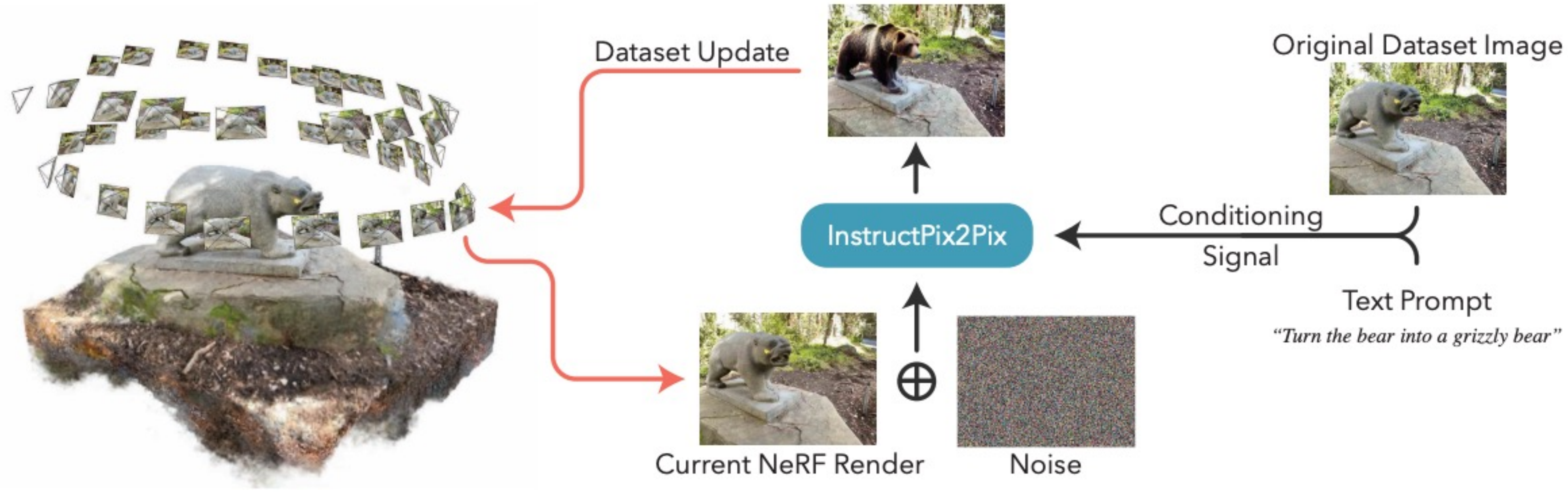
DreamFusion: Text-to-3D using 2D Diffusion [Poole et al., 2022]

Instruct NeRF2NeRF



Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions [Haque et al., 2023]

Instruct NeRF2NeRF



FOR loop

- Step 1. Render a view using existing NeRF
- Step 2. Use InstructPix2Pix to produce output images
- Step 3. Update NeRF parameters with the generated result from Step 2

InstructPix2pix: image-conditional diffusion model (https://www.timothybrooks.com/instruct-pix2pix/)

Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions [Haque et al., 2023]

Video Synthesis and Editing

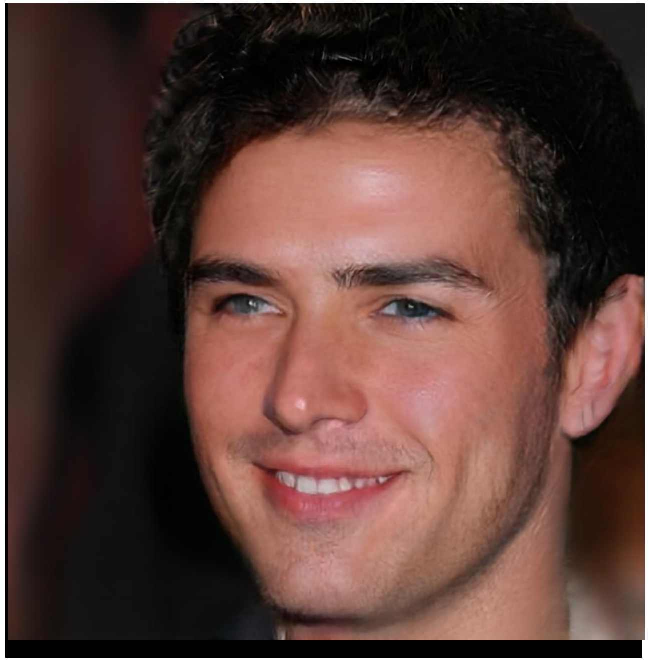
Jun-Yan Zhu

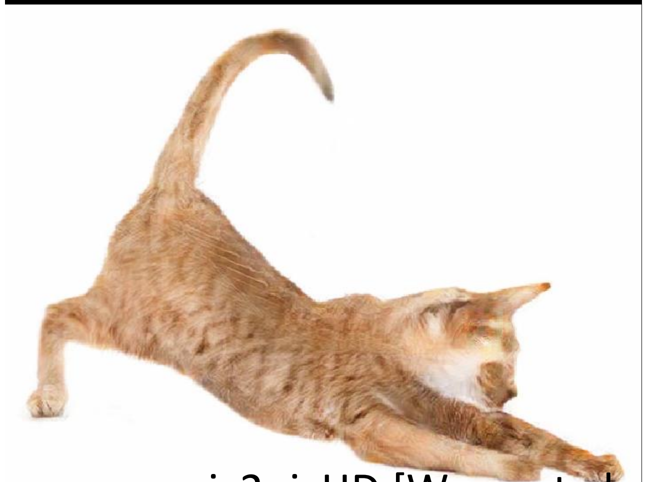
16-726 Learning-based Image Synthesis, Spring 2023

Many slides adopted from Alexei A. Efros, Ting-Chun Wang, Ming-Yu Liu, Caroline Chan, Tinghui Zhou, Shiry Ginosar
See Stanford's course Computational Video Manipulation (https://magrawala.github.io/cs448v-sp19/) for more details

Video © Wang et al., NeurIPS 2018. vid2vid

Image Editing and Synthesis



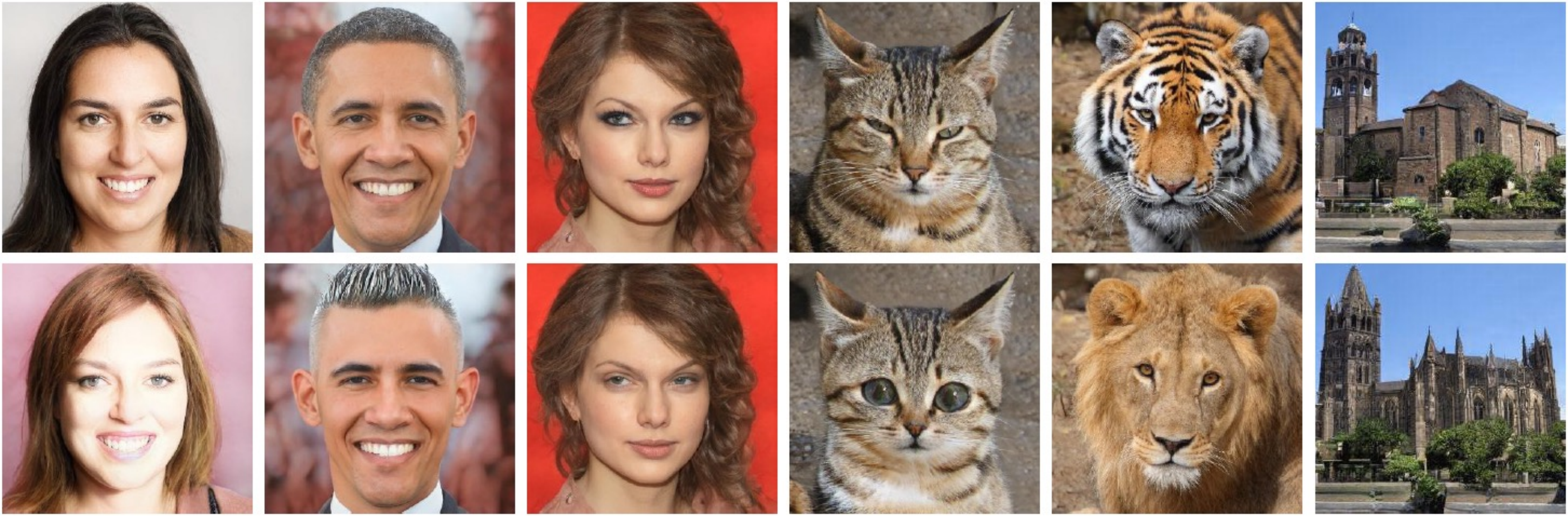


pix2pixHD [Wang et al., CVPR 2018]

Image Editing and Synthesis

CycleGAN [Zhu et al., ICCV 2017]

Image Editing and Synthesis



“Emma Stone” “Mohawk hairstyle” “Without makeup” “Cute cat” “Lion” “Gothic church”

StyleCLIP [Patashnik et al., 2021]

Images vs. Videos

- What are the major differences?
  - 2D vs. 3D
  - Spatial vs. spatial + temporal
- Differences between 3D vs. videos?
  - (x, y, z) vs. (x, y, t)
- Why are videos much more challenging?
  - Understanding motions/actions/intensions.
  - Long-term dependence.
  - Hard to annotate.
  - Computationally-expensive (e.g., memory, training time)

Why not apply it to video?

Someone gave you an image synthesis model

asked you to build a video synthesis application

How to Generalize to Videos

- Idea 1: Frame-by-Frame

Frame-by-Frame Result (pix2pixHD)

Frame-by-Frame Result (CycleGAN)

How to Generalize to Videos

- Idea 1: Frame-by-Frame
  - Temporal inconsistency (flicking, color drift)
- Idea 2: Video as 3D data (height x width x time)

Case Study: 3D Poisson blending

Spatial-temporal Constraints

$$
\begin{aligned}
F(\nabla I, G) &= \|\nabla I - G\|^2 \\
&= \left(\frac{\partial I}{\partial x} - G_x\right)^2 + \left(\frac{\partial I}{\partial y} - G_y\right)^2 + \left(\frac{\partial I}{\partial t} - G_t\right)^2
\end{aligned}
$$

Output Image

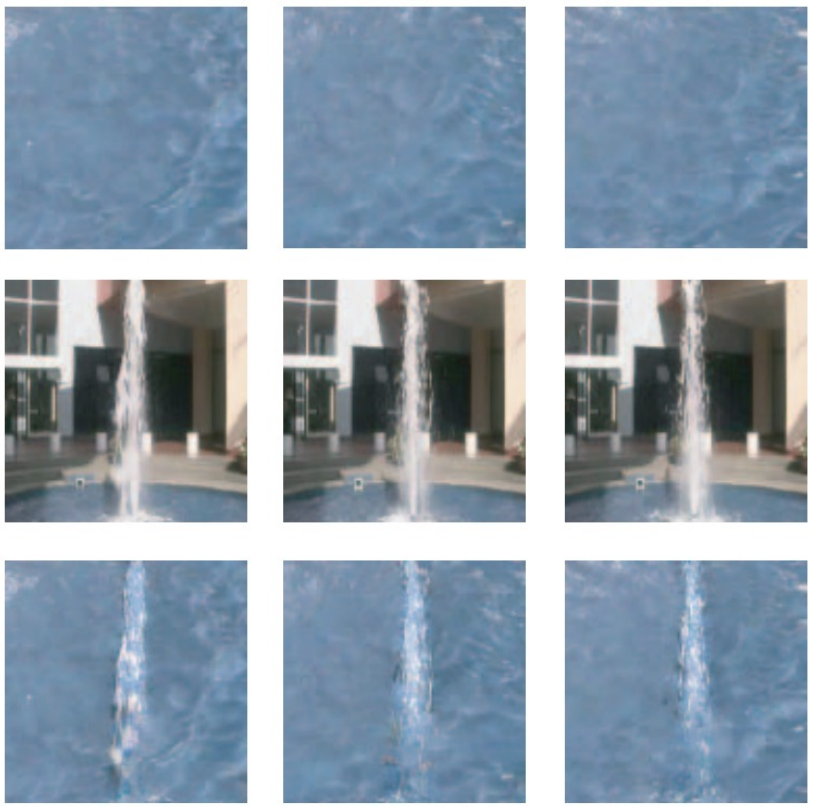
Guidance Gradient

Time

Background

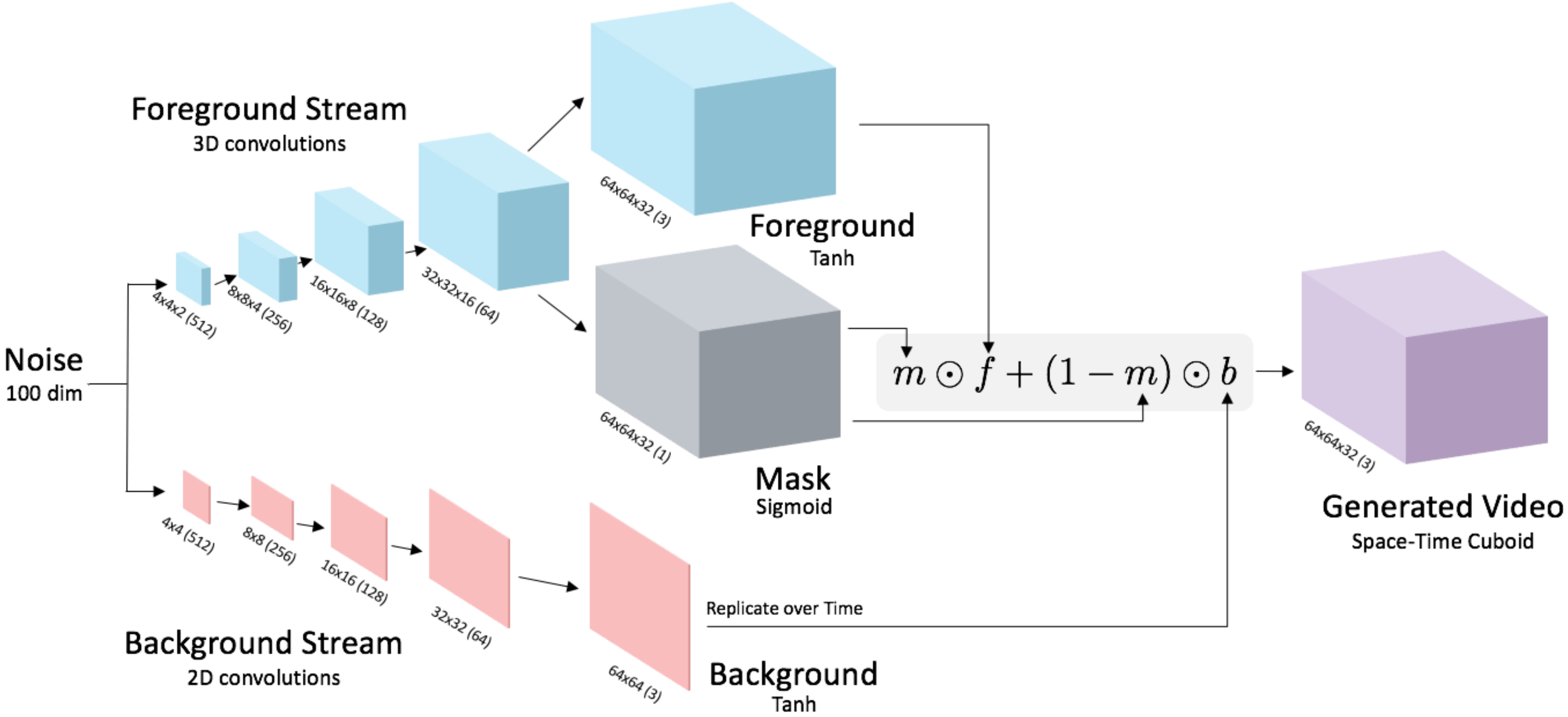
Foreground

Output



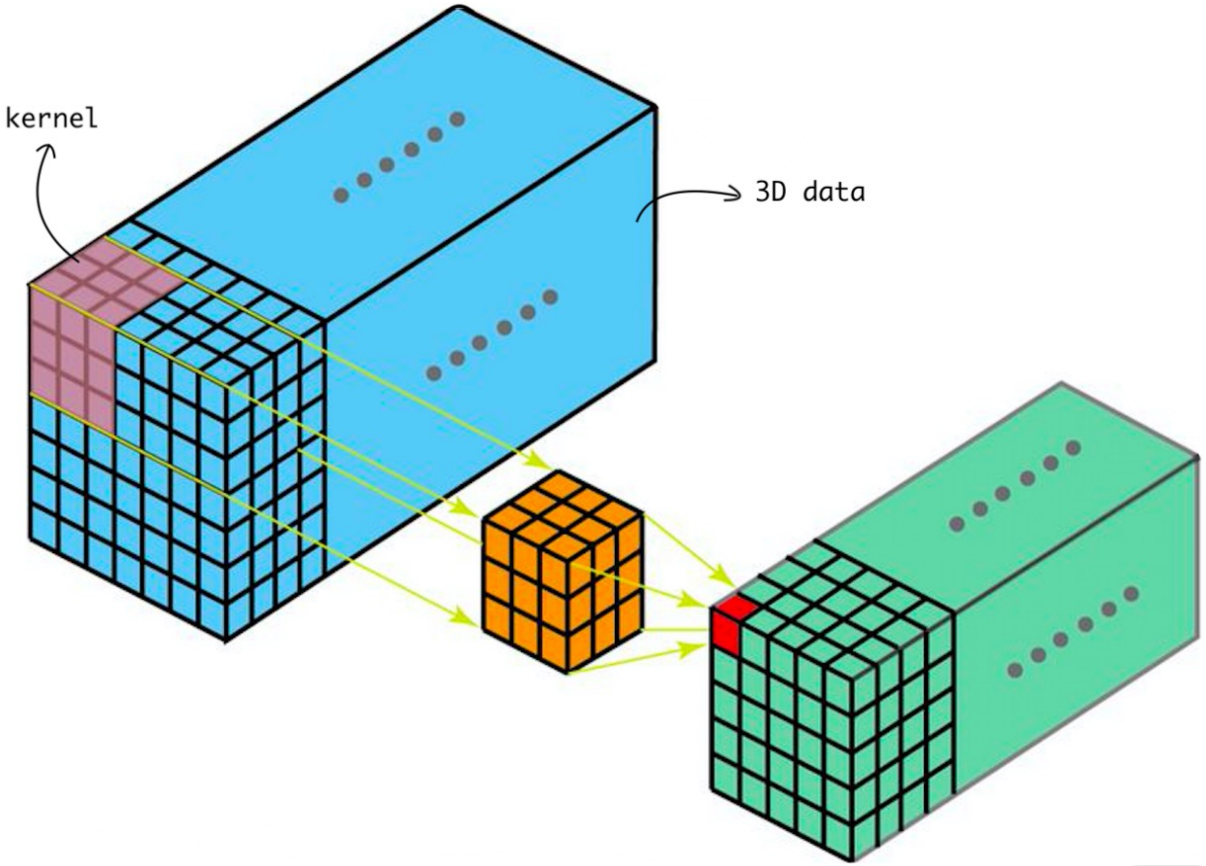
H. Wang, R. Raskar, and N. Ahuja, "Seamless video editing," in ICPR 2004

Case Study: Video GANs



Vondrick et al., NeurIPS 2016

Recap: 3D Conv



Easy to implement:
- Replace 2D by 3D in your code

e.g., Conv2D -> Conv3D
ConvTranspose2d->ConvTranspose3d
MaxPool2d -> MaxPool3d

CLASS `torch.nn.Conv3d(in_channels, out_channels, kernel_size, stride=1, padding=0, dilation=1, groups=1, bias=True, padding_mode='zeros', device=None, dtype=None)` [SOURCE]

Case Study: Video GANs

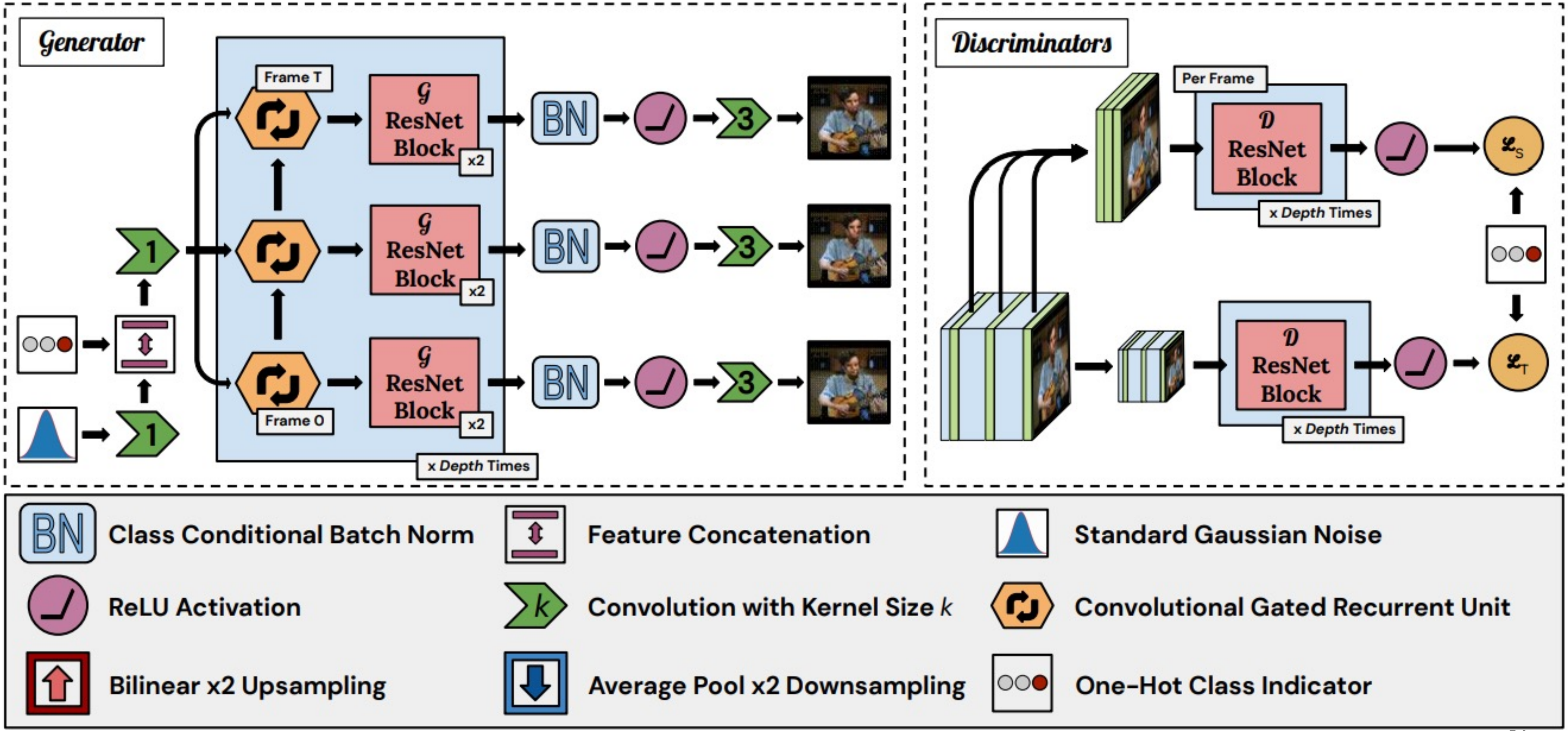
Vondrick et al., NeurIPS 2016

Case Study: DVDGANs

BigGAN-based generator + Spatial Discriminator + Temporal Discriminator

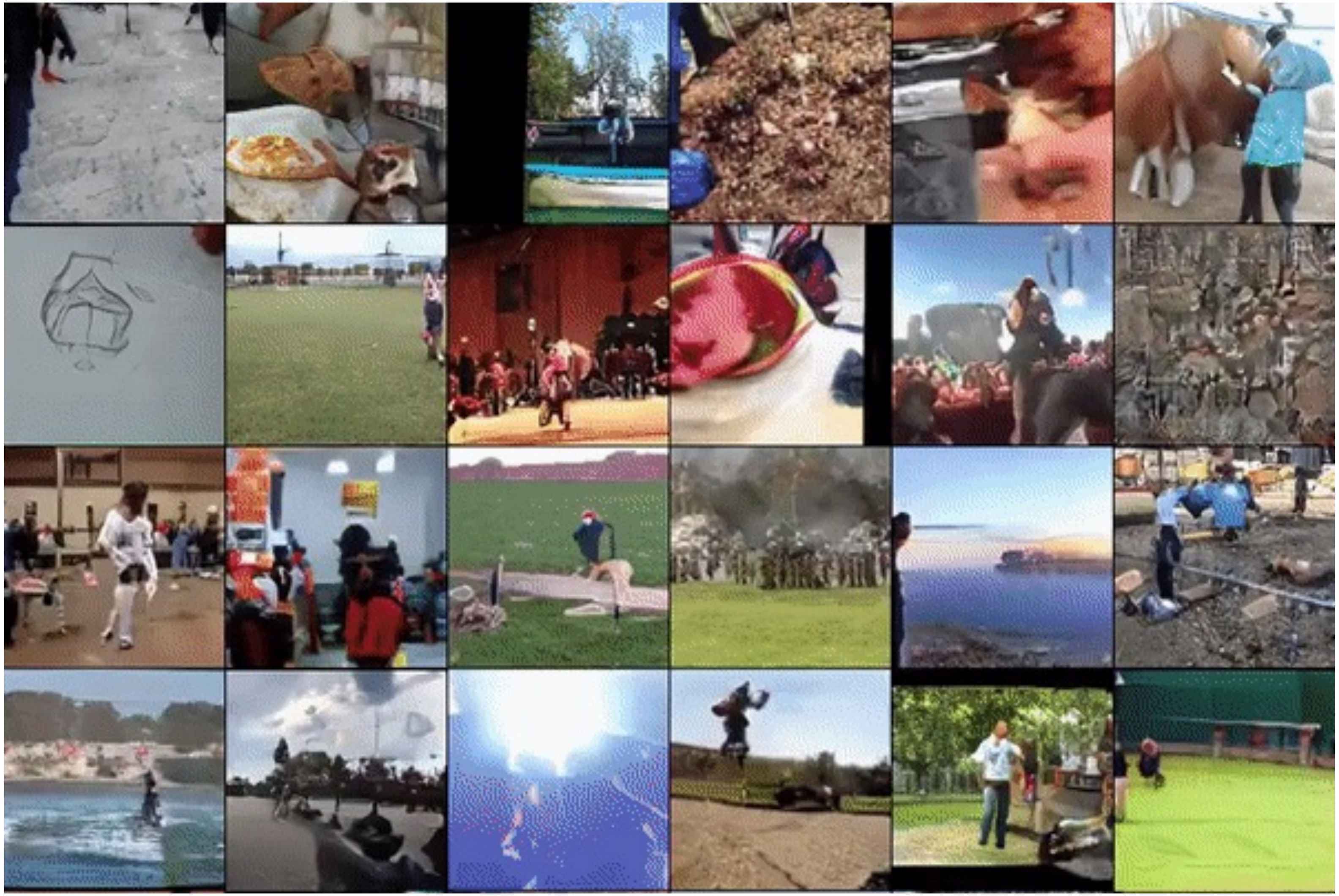
Clark et all, arXiv 2019

Case Study: DVDGANs



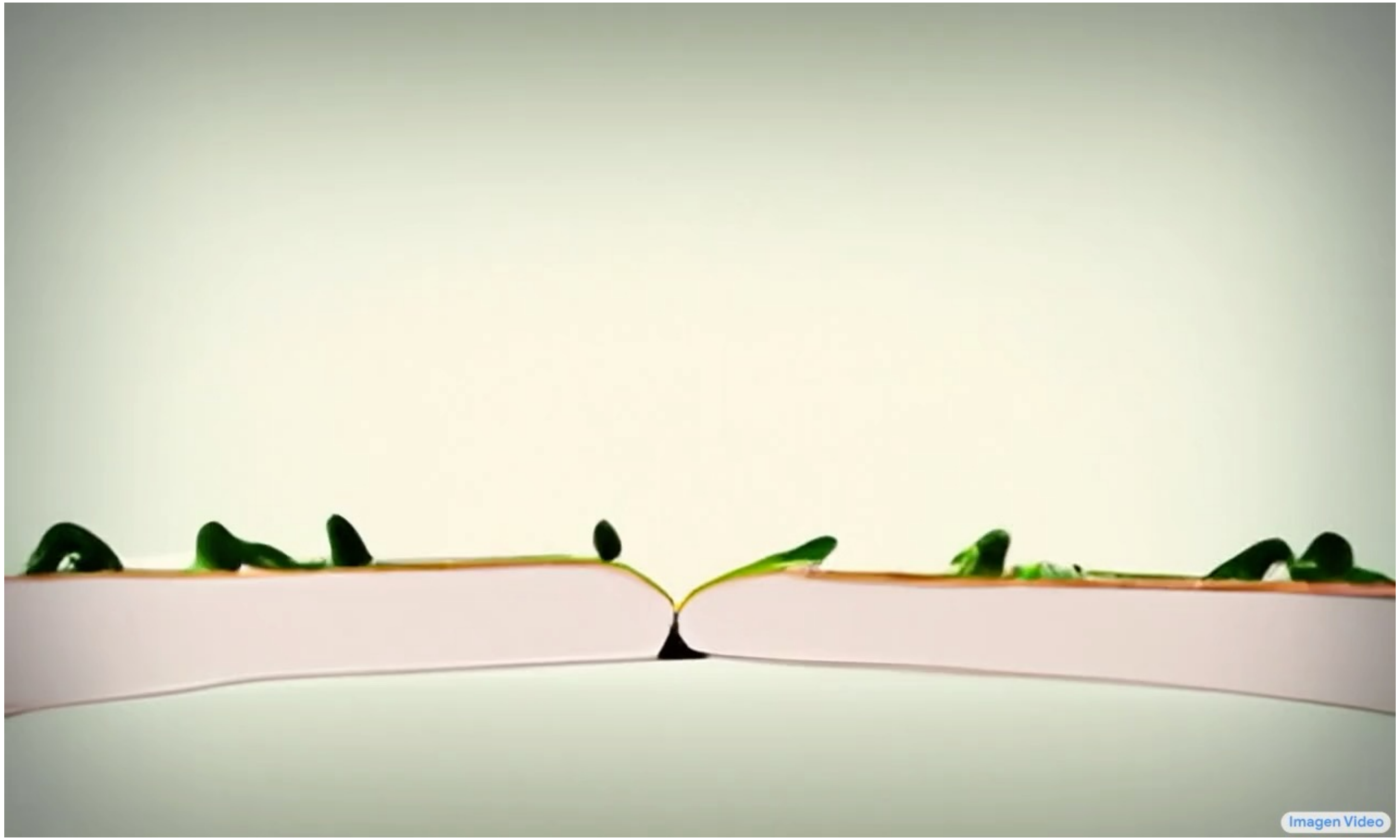
Clark et all, arXiv 2019

Case Study: DVDGANs



Clark et all, arXiv 2019

Case Study: Imagen Video



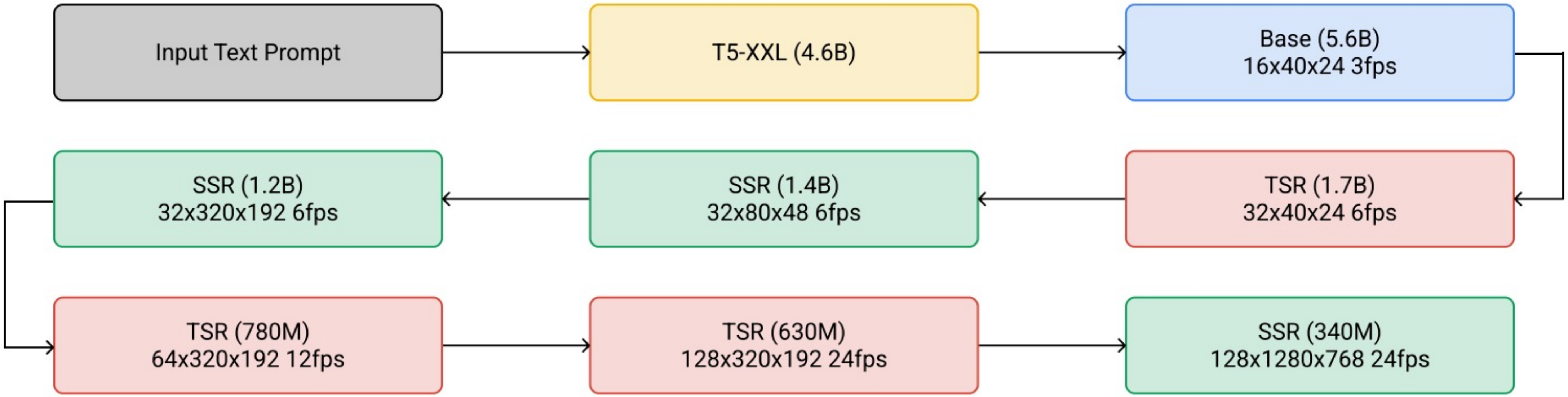
Case Study: Imagen Video

An astronaut riding a horse

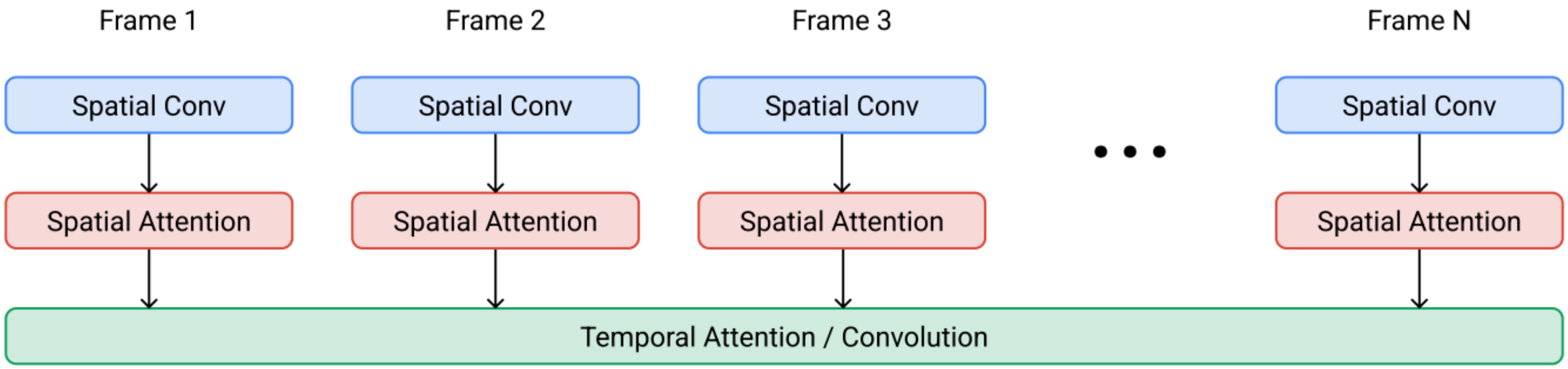
Case Study: Imagen Video

Wooden figure surfing on a surfboard in space

Case Study: Imagen Video



Case Study: Imagen Video



How to Generalize to Videos

- Idea 1: Frame-by-Frame
  - Temporal inconsistency (flicking, color drift)
- Idea 2: Video as 3D data (height x width x time)
  - memory-intensive and time-consuming
  - only work for a short video at low resolution
- Idea 3: recurrent (autoregressive) synthesis
  - Generate 1^{st} frame, generate 2^{nd} frame based on 1^{st} one, …
  - Using optical flow (optional): warp 1^{st} frame to 2^{nd} frame.

Text Synthesis

- [Shannon,'48] proposed a way to generate English-looking text using N-grams:
  - Assume a generalized Markov model
  - Use a large text to compute prob. distributions of each letter given N-1 previous letters
  - Starting from a seed repeatedly sample this Markov chain to generate new letters
  - Also works for whole words

WE NEED TO EAT CAKE

Slides from Alyosha Efros

Case Study: Video Textures

Still image

Loopy video

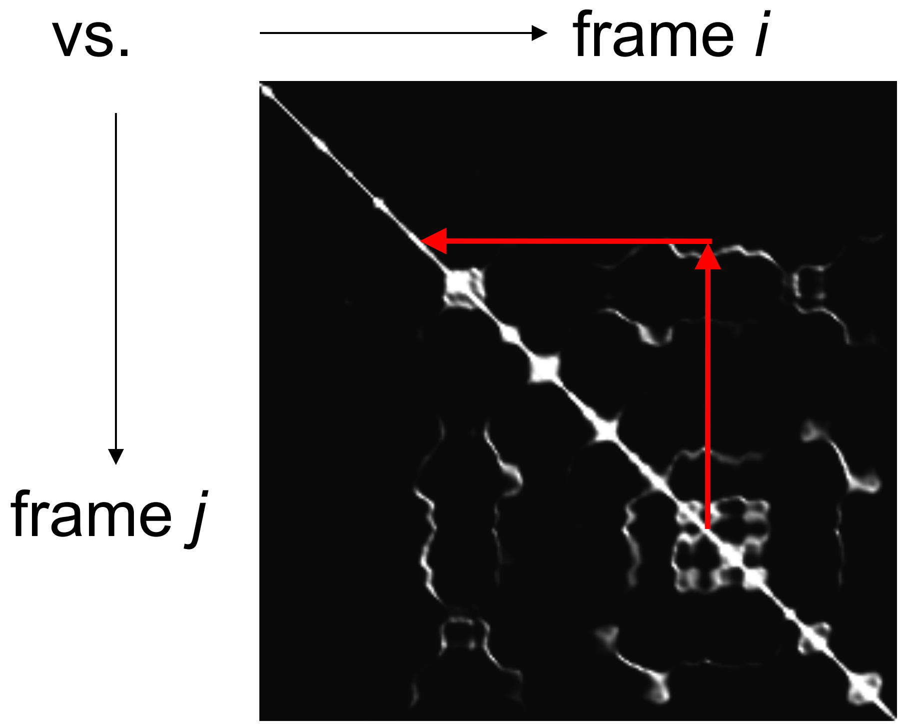
[Schödl et al., SIGGRAPH 2000]

Case Study: Video Textures

[Schödl et al., SIGGRAPH 2000]

Finding good transitions

- Compute L_2 distance $D_{i,\,j}$ between all frames



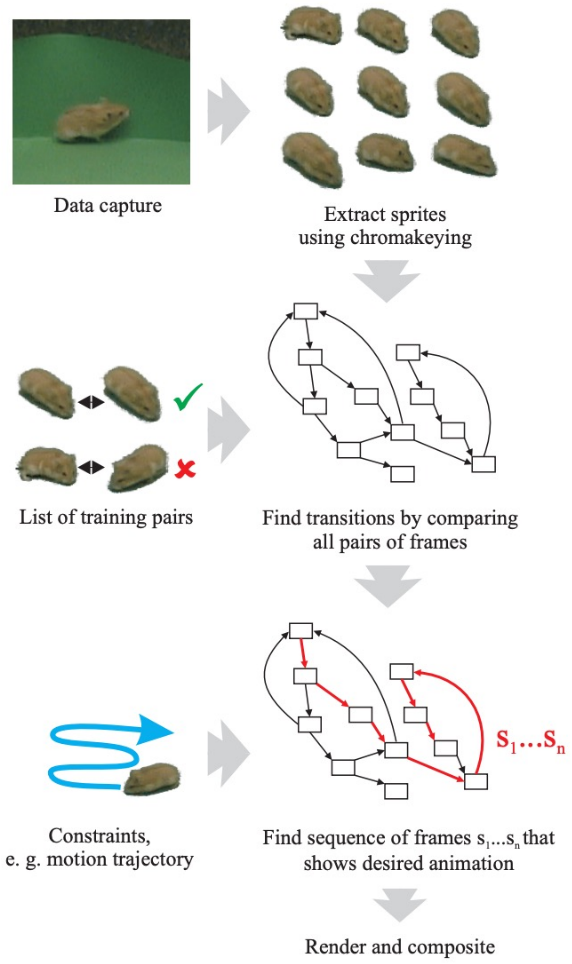
Similar frames make good transitions

[Schödl et al., SIGGRAPH 2000]

Case Study: Video Textures

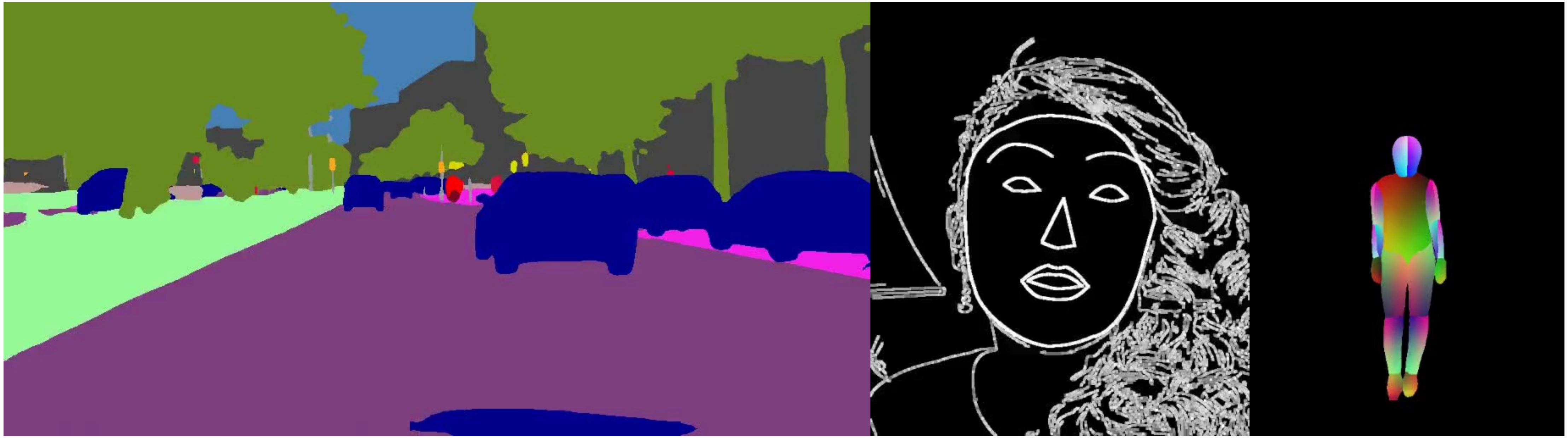
[Schödl et al., SIGGRAPH 2000]

Case Study: Controlled Animation of Video Sprites



[Arno Schodl and Irfan A. Essa, SIGGRAPH 2002]

Case Study: Video-to-Video Translation



T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz, B. Catanzaro,
"Video-to-Video Synthesis," NeurIPS 2018.
https://github.com/NVIDIA/vid2vid

Previous Work: Frame-by-Frame Result

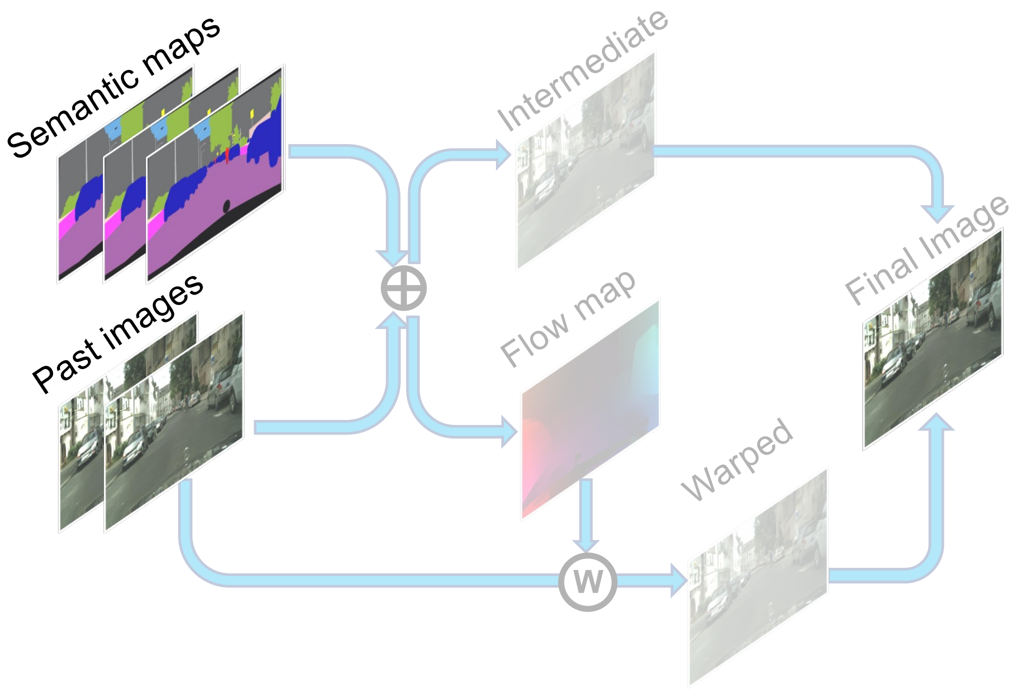
vid2vid

- Sequential generator
- Multi-scale temporal discriminator
- Spatio-temporal progressive training procedure

Vid2vid [Wang et al., NeurIPS 2018]

vid2vid

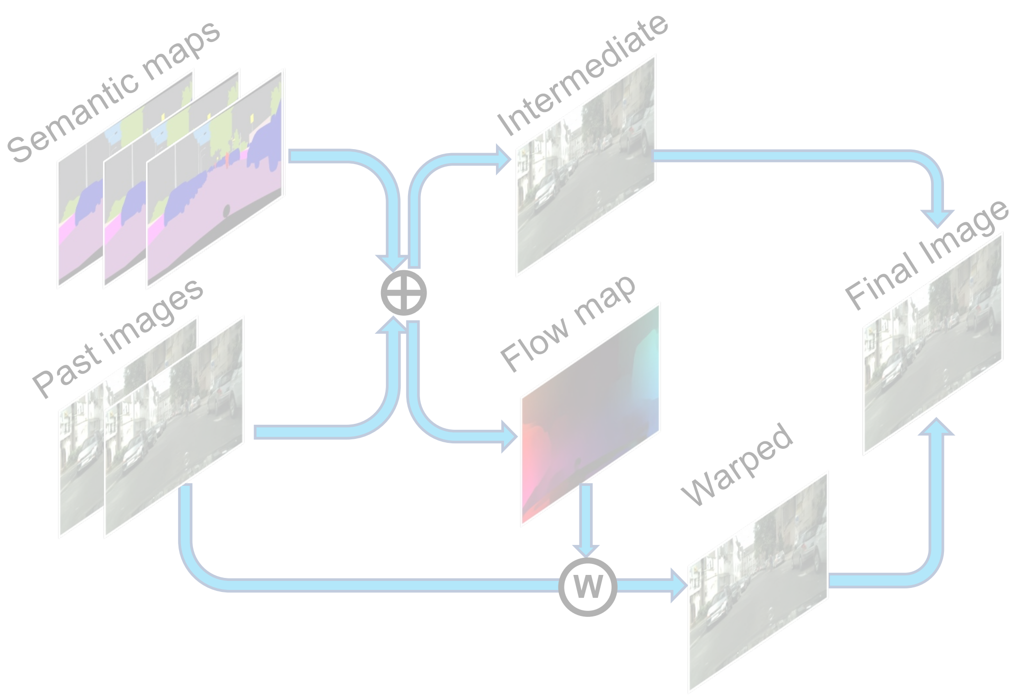
Sequential Generator



Vid2vid [Wang et al., NeurIPS 2018]

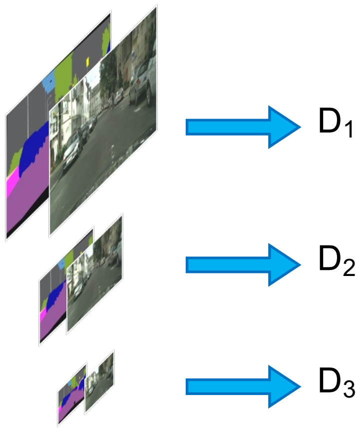
vid2vid

Sequential Generator

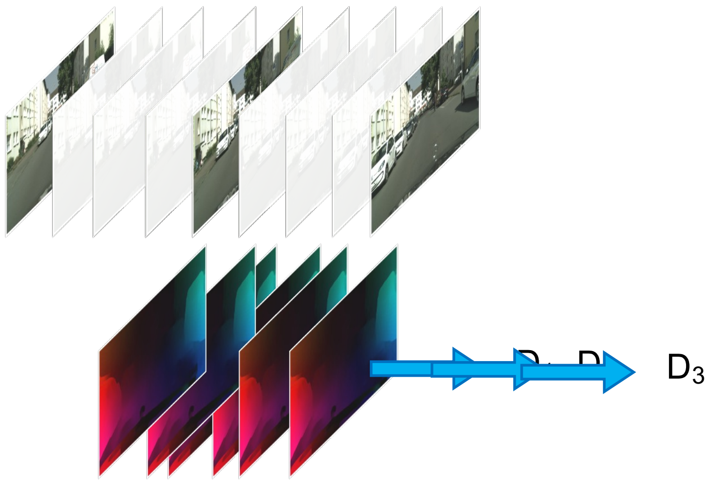


Multi-scale Discriminators

Image Discriminator



Video Discriminator

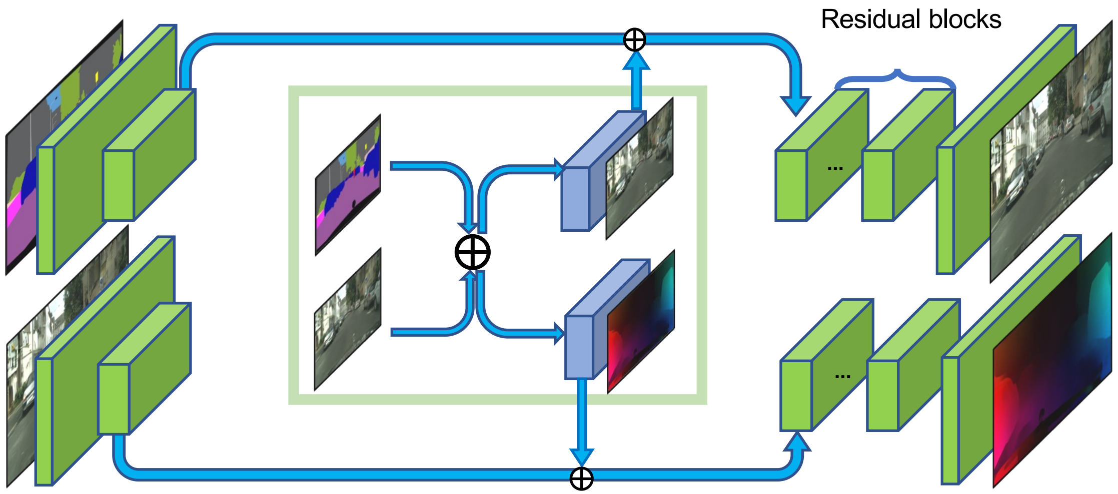


Vid2vid [Wang et al., NeurIPS 2018]

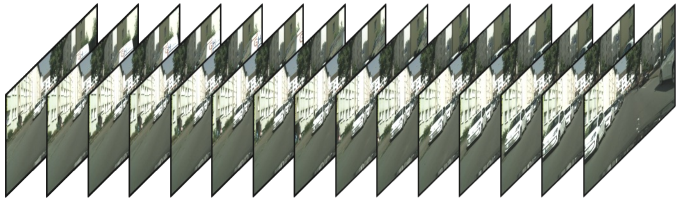
vid2vid

Spatio-temporally Progressive Training

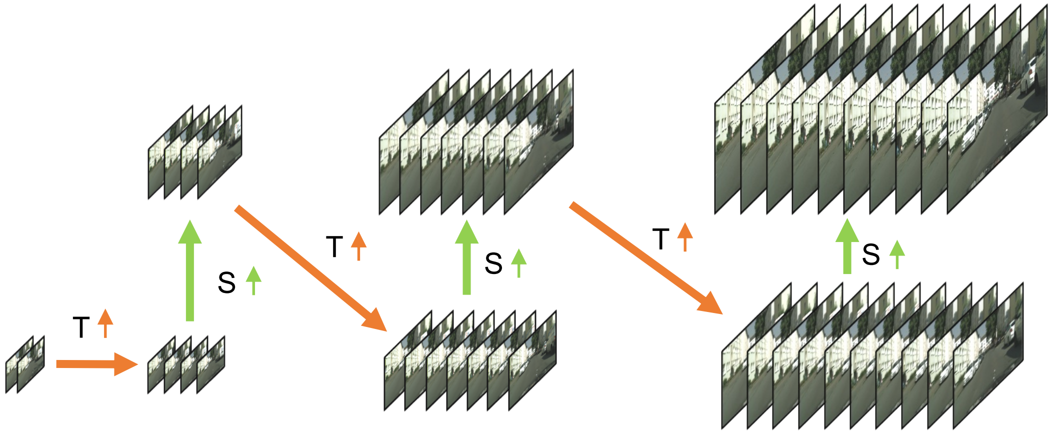
Spatially progressive



Temporally progressive



Alternating training



Vid2vid [Wang et al., NeurIPS 2018]

vid2vid Results

- Semantic → Street view scenes
- Edges → Human faces
- Poses → Human bodies

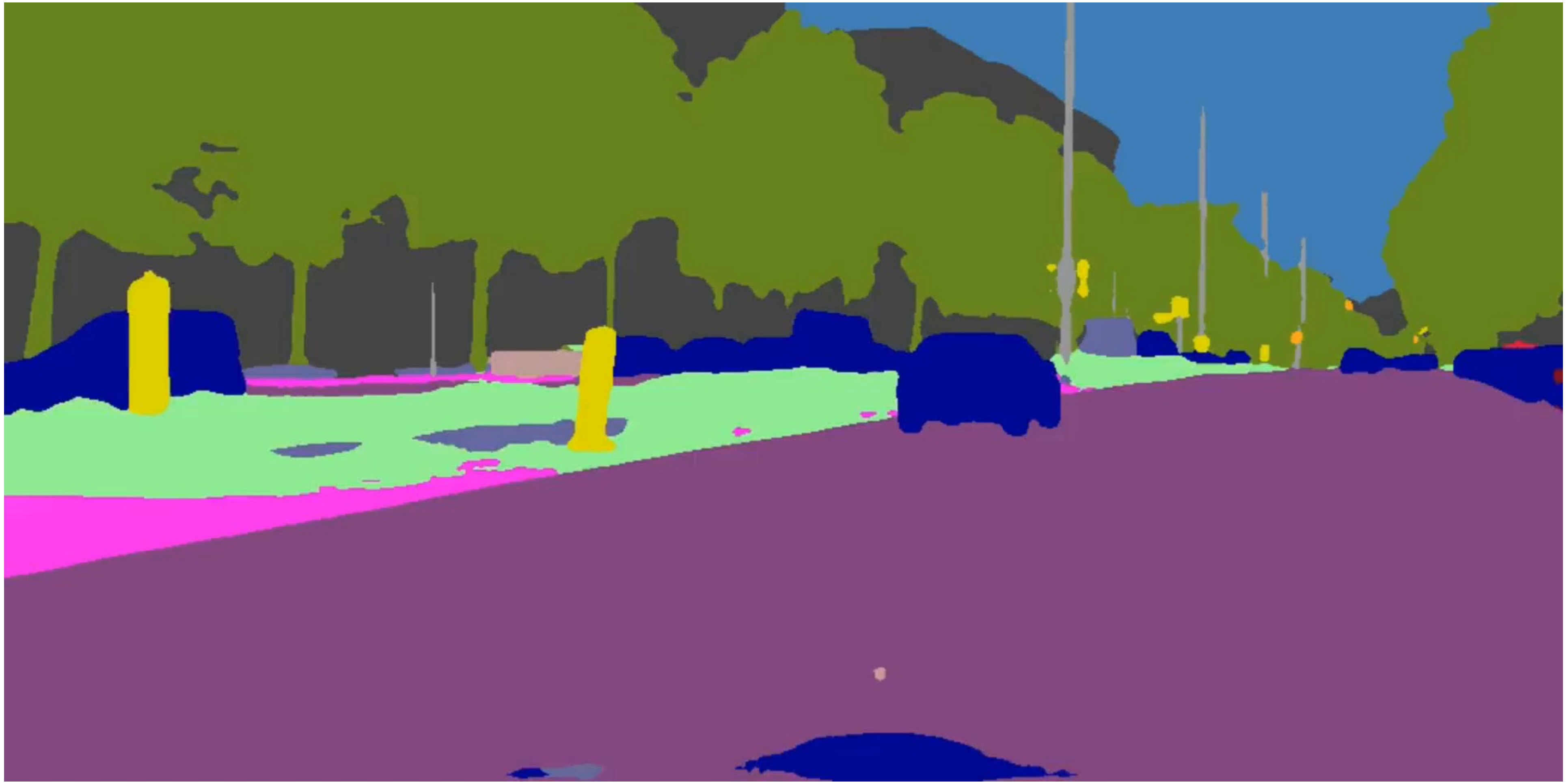
Vid2vid [Wang et al., NeurIPS 2018]

vid2vid Results

- Semantic → Street view scenes

Vid2vid [Wang et al., NeurIPS 2018]

Street View: Cityscapes

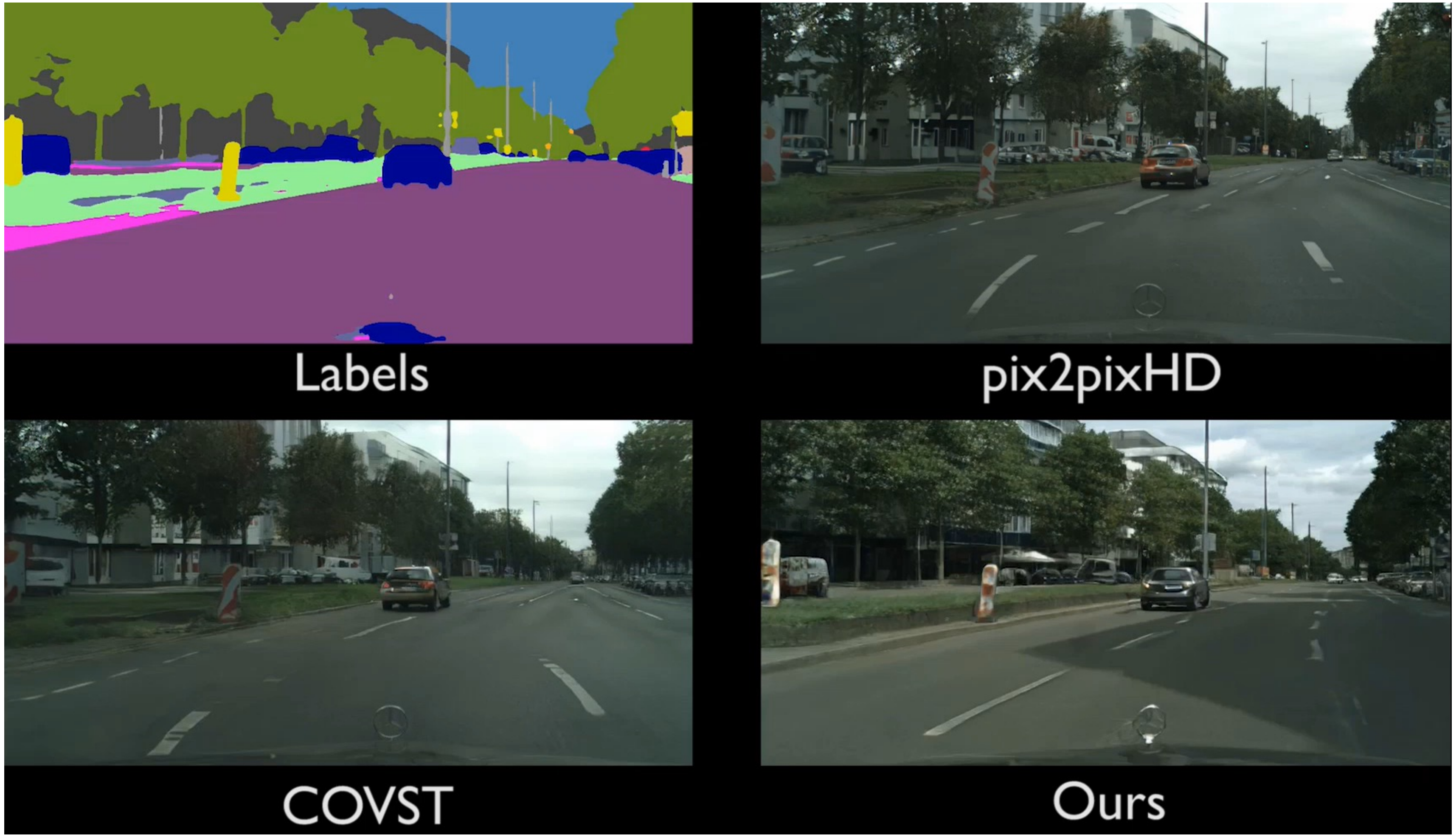


Vid2vid [Wang et al., NeurIPS 2018]

Street View: Cityscapes

Vid2vid [Wang et al., NeurIPS 2018]

Street View: Cityscapes



Vid2vid [Wang et al., NeurIPS 2018]

Street View: Boston


Vid2vid [Wang et al., NeurIPS 2018]

Street View: NYC

Vid2vid [Wang et al., NeurIPS 2018]

Results

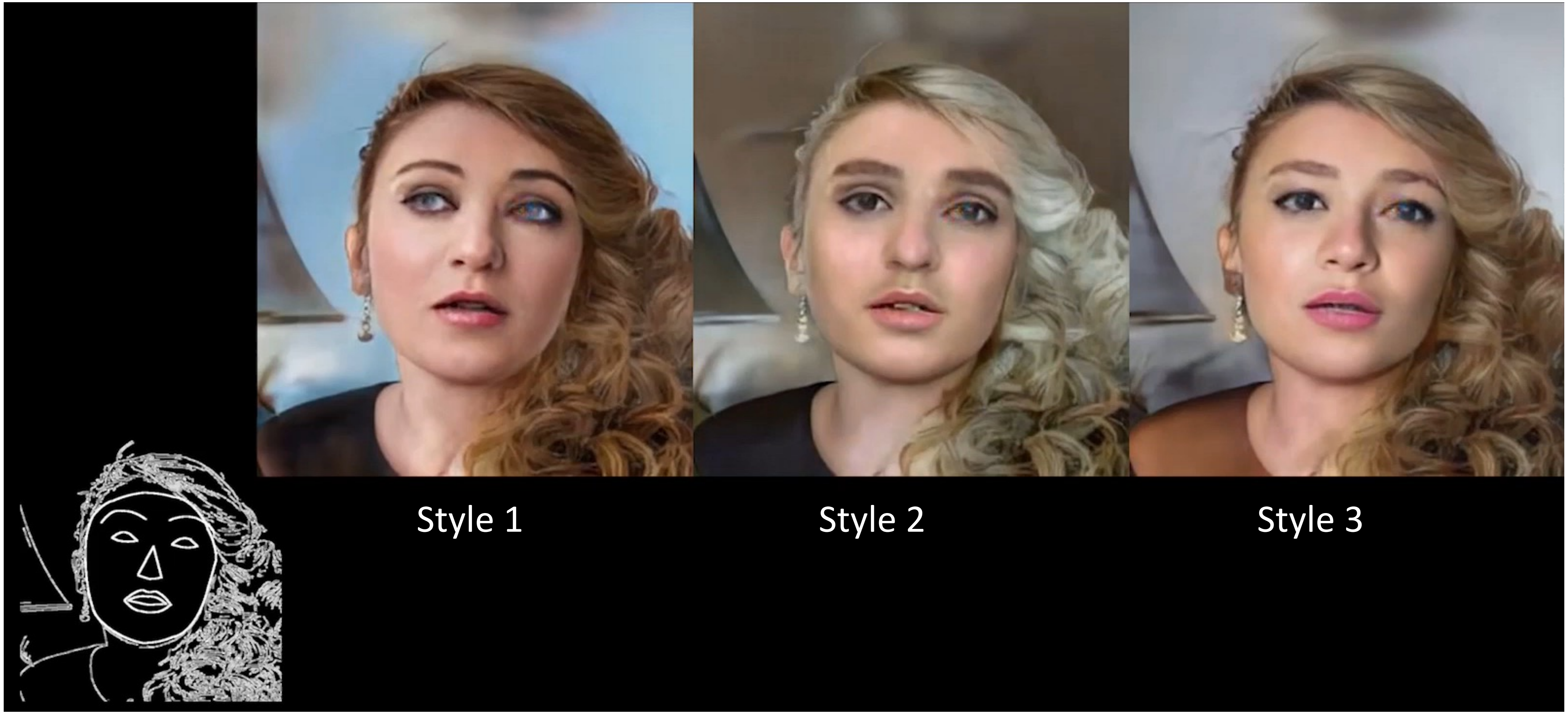
- Edges → Human faces

Face Swapping (face → edge → face)

input edges output

Vid2vid [Wang et al., NeurIPS 2018]

Multi-modal Edge → Face



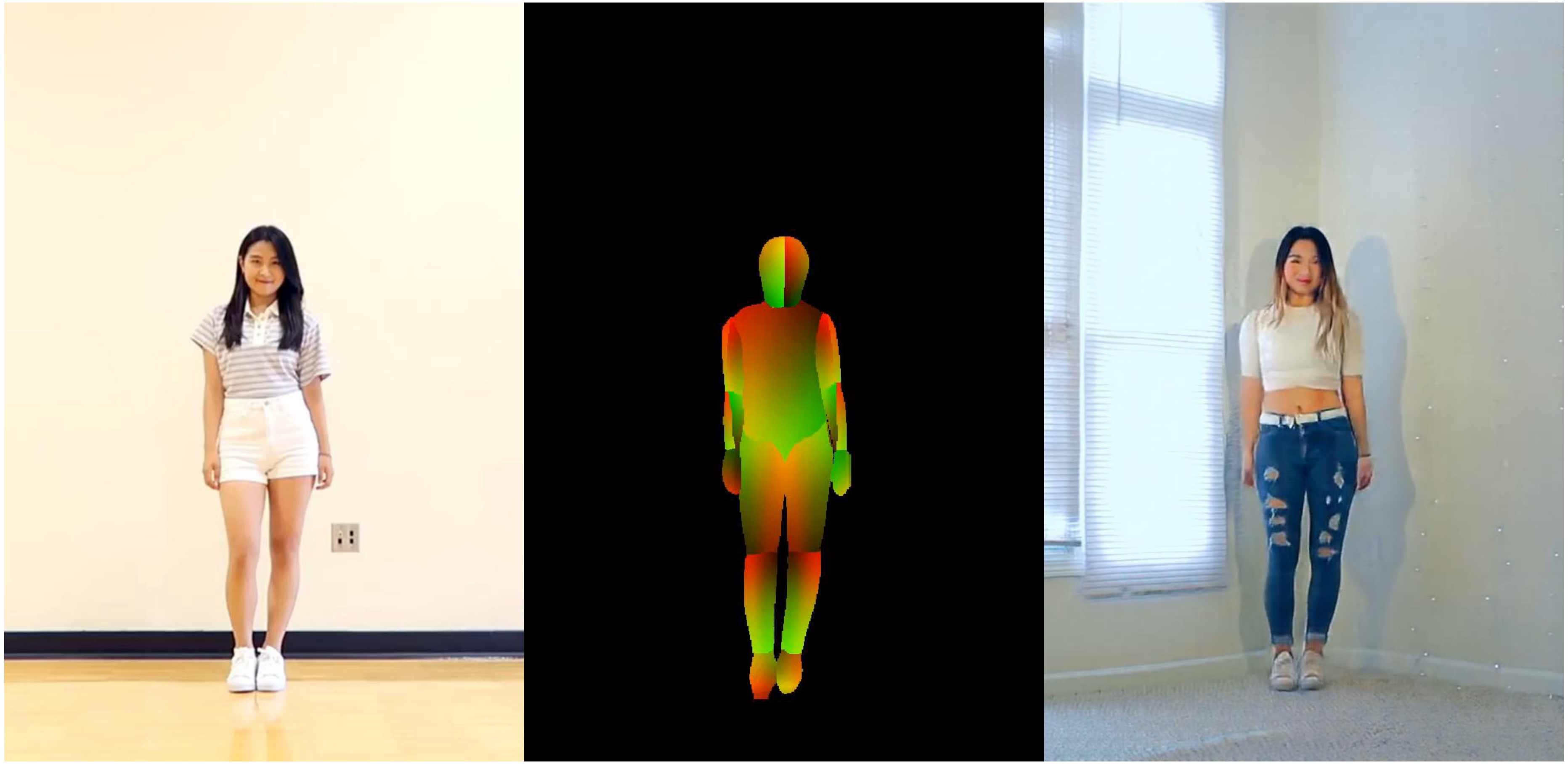
Results

- Poses → Human bodies

Motion Transfer (body → pose → body)

input poses output

Motion Transfer (body → pose → body)



input poses output

Motion Transfer (body → pose → body)

input poses output

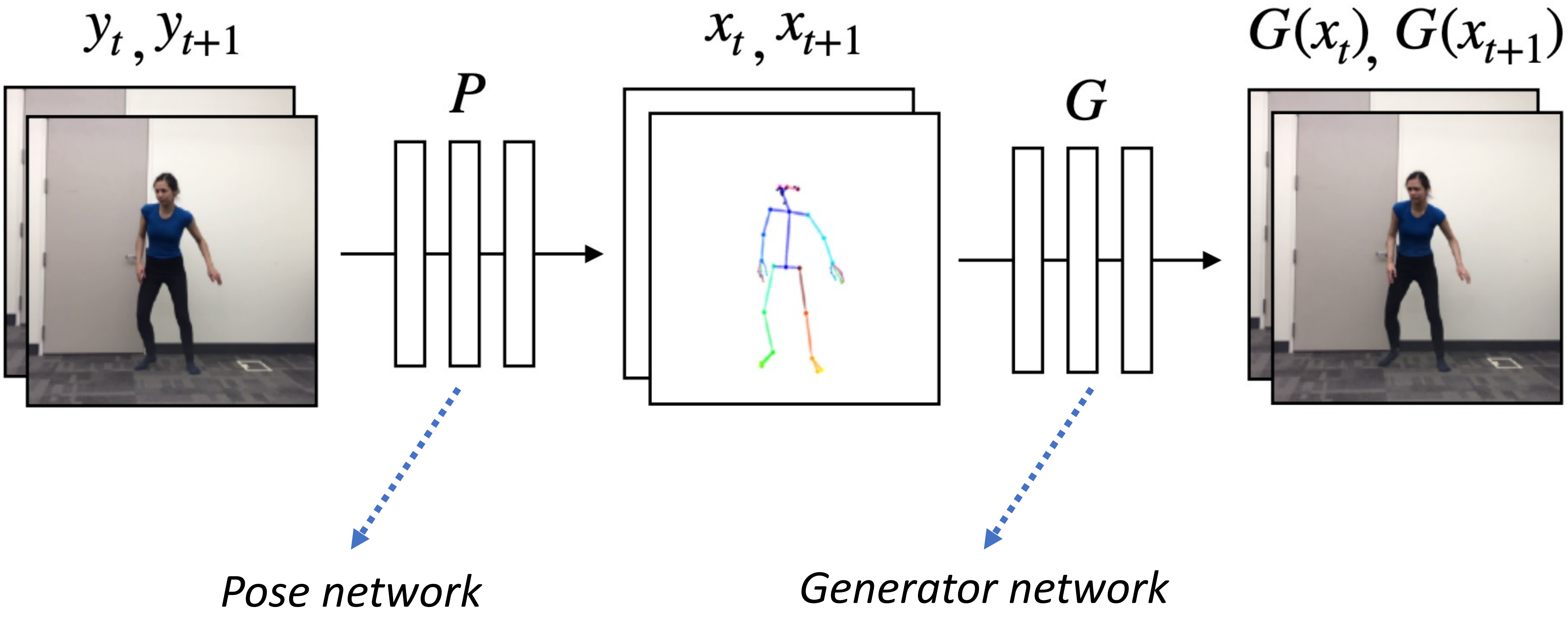
Motion Transfer (body → pose → body)

input poses output

More Dancing …

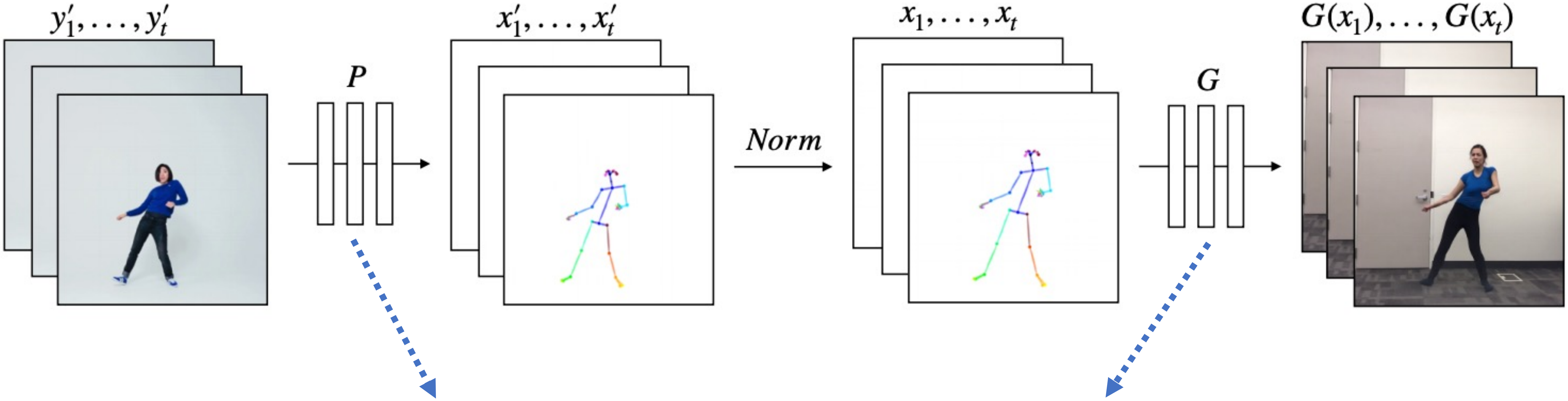
Everybody Dance Now [Chan et al., ICCV 2019]

Pose-guided Synthesis



Everybody Dance Now [Chan et al., ICCV 2019]

Transfer Phase



Pose network

Generator network

Everybody Dance Now [Chan et al., ICCV 2019]

Everybody Dance Now [Chan et al., ICCV 2019]

ML-based Rendering

Traditional graphics

Geometry, texture, lighting

Machine learning graphics

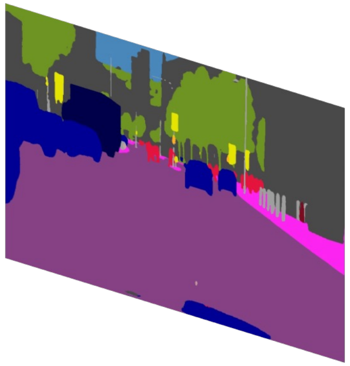
Data

Vid2vid [Wang et al., NeurIPS 2018]

vid2vid Extensions: Interactive Graphics

User control

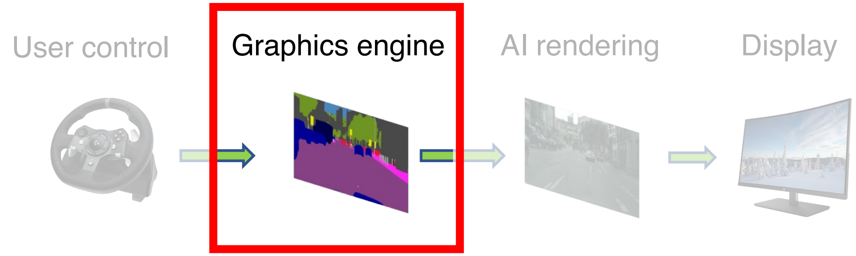
Graphics engine



AI rendering

Display

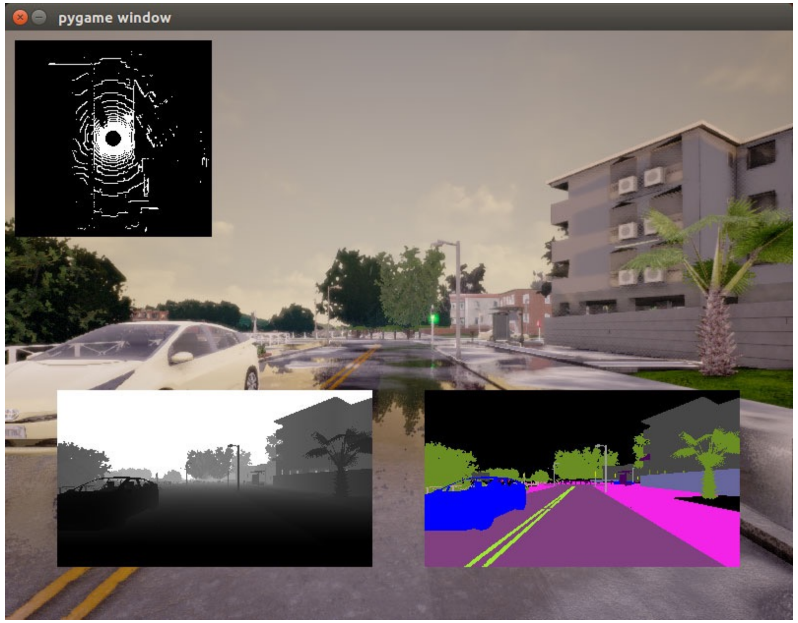
Graphics Engine: CARLA



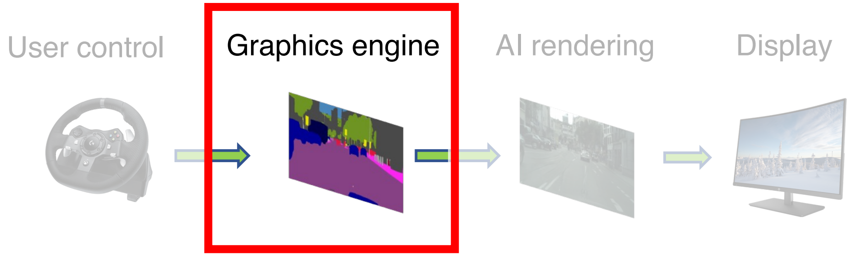
DATA

PROTOBUF messages

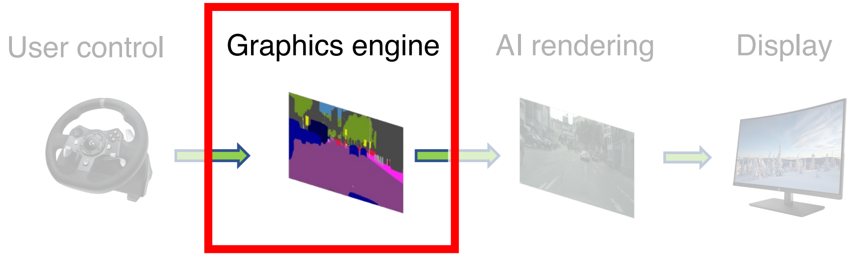
CONTROL

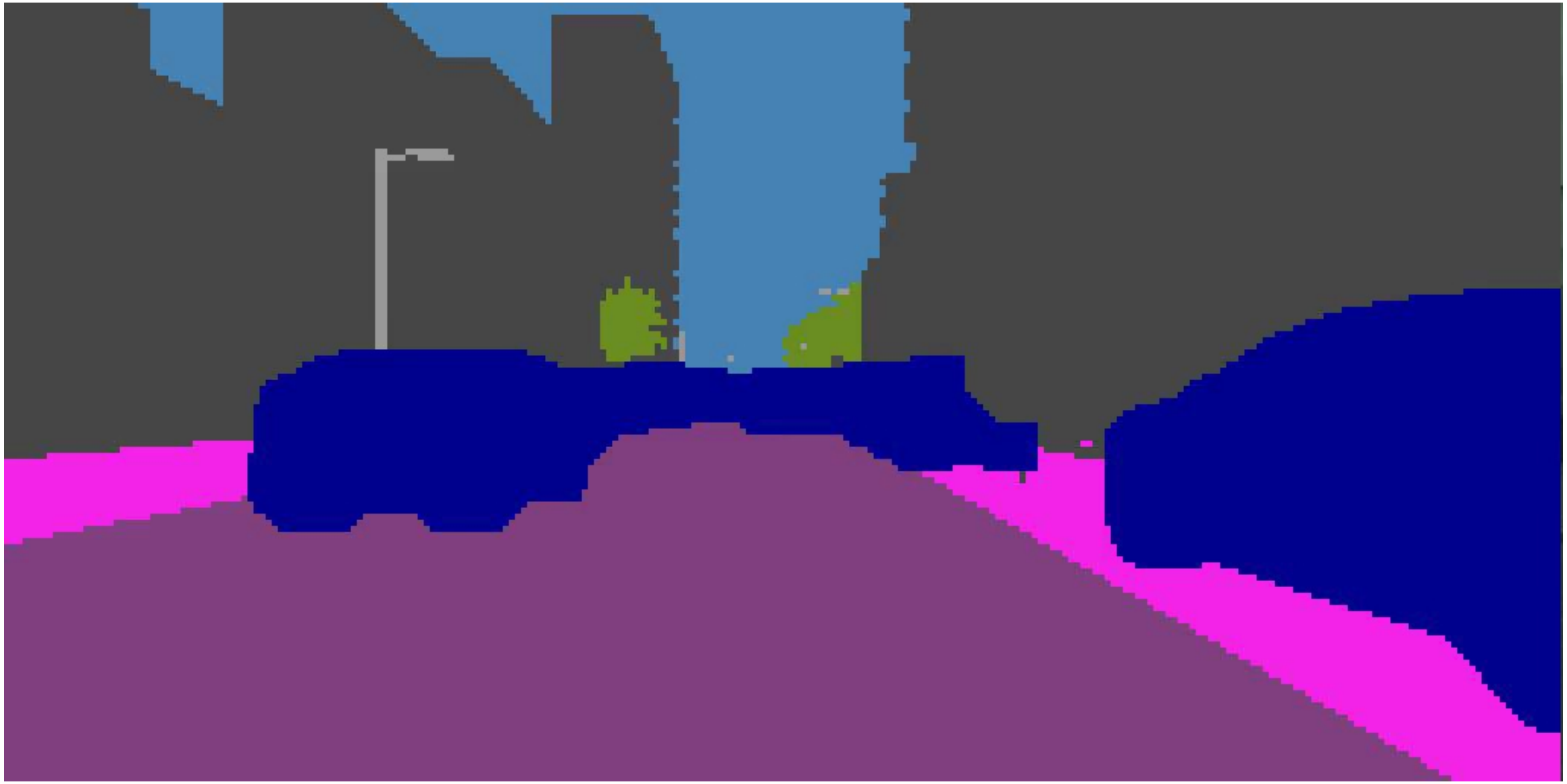


Original CARLA Sequence

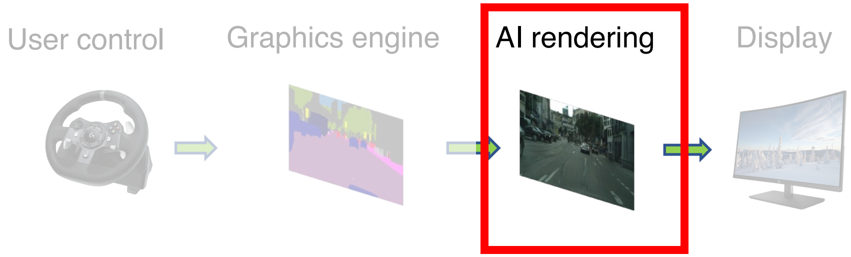


CARLA Semantic Maps

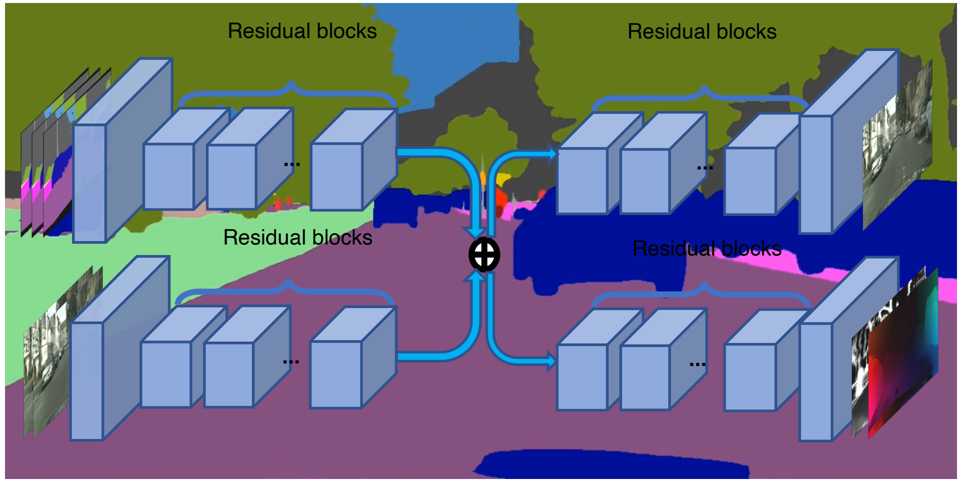


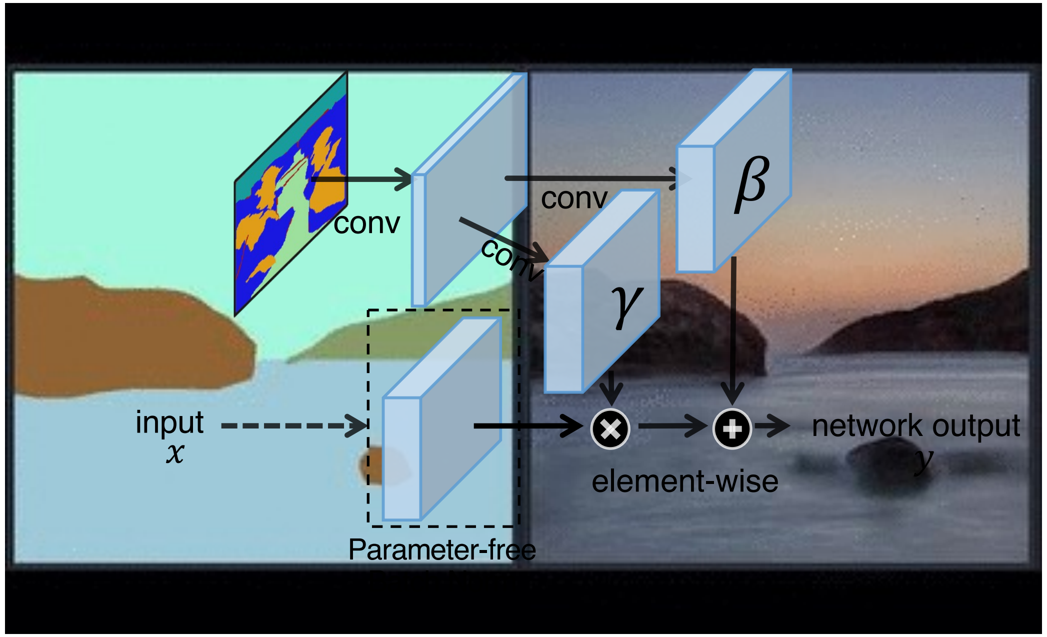


Methodology

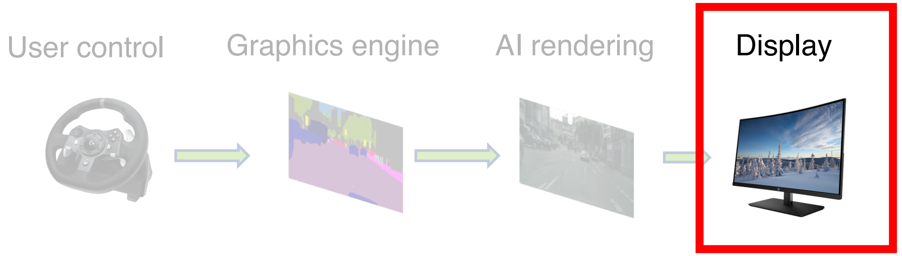


- Combine vid2vid with SPADE





Demo Result



Driving Game

vid2game by FAIR

O. Gafni, L. Wolf, Y. Taigman. "Vid2Game: Controllable Characters Extracted from Real-World Videos," 2019

Thank You!

16-726, Spring 2023

https://learning-image-synthesis.github.io/

Video © Background Matting [Sengupta et al., CVPR 2020]