Video Synthesis and Editing

Jun-Yan Zhu

16-726 Learning-based Image Synthesis, Spring 2022

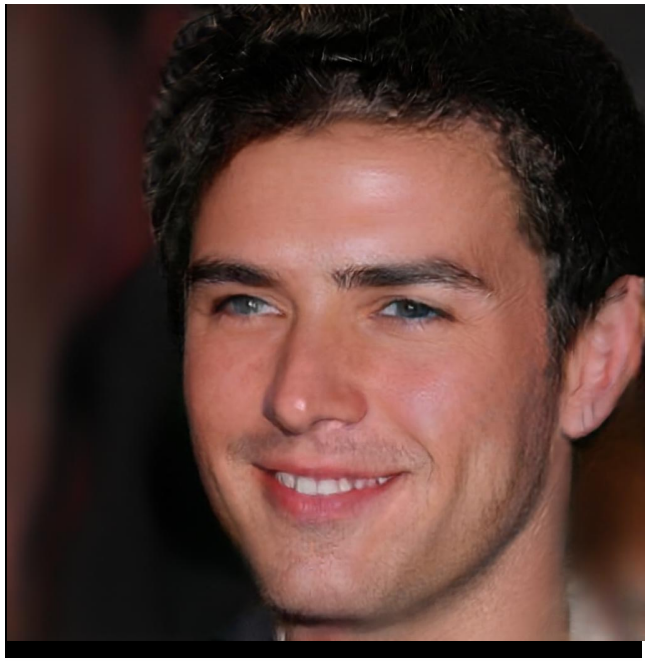
Many slides adopted from Alexei A. Efros, Ting-Chun Wang, Ming-Yu Liu, Caroline Chan, Tinghui Zhou, Shiry Ginosar
See Stanford's course Computational Video Manipulation (https://magrawala.github.io/cs448v-sp19/) for more details

Video © Wang et al., NeurIPS 2018. vid2vid

Logistics

- HW5 award on Wed.
- 5 min (including Q & A) x 23 groups / 60 min ~ 2 hours
- Strict timing.
- Preliminary results.
- Common failure cases:
  - You didn’t show up.
  - You gave a 10 min talk.
  - You spent 4.5 min on related work.
  - We cannot open your link.

Image Editing and Synthesis

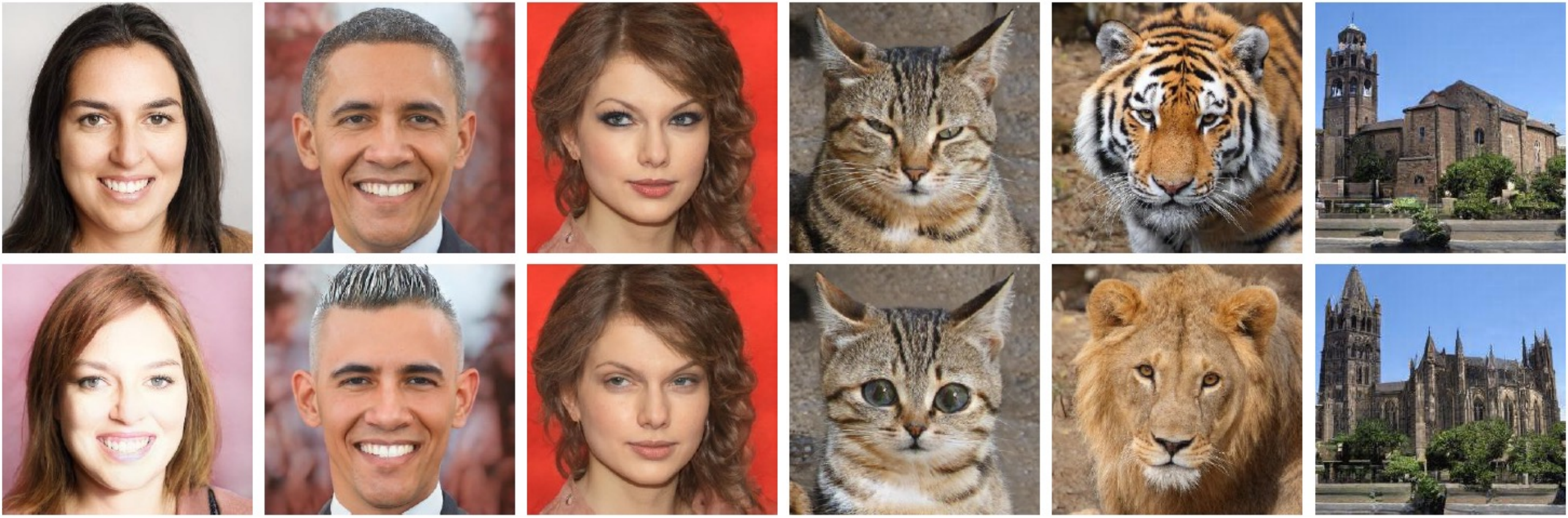


pix2pixHD [Wang et al., CVPR 2018]

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)

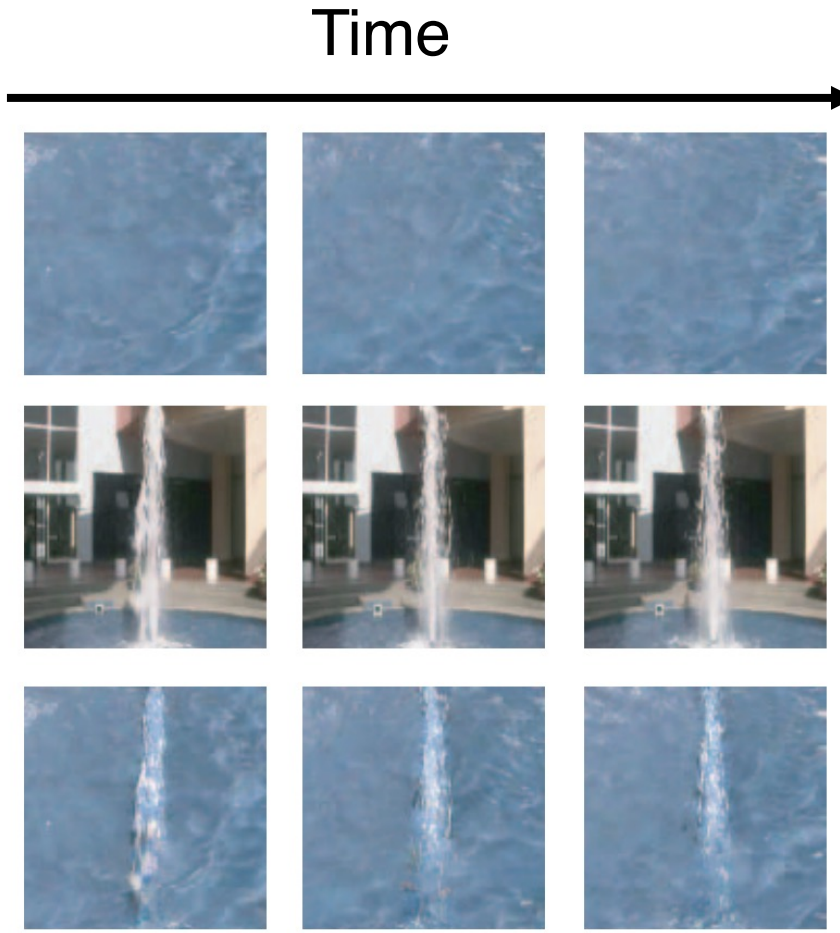
Case Study: 3D Poisson blending

Spatial-temporal Constraints

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

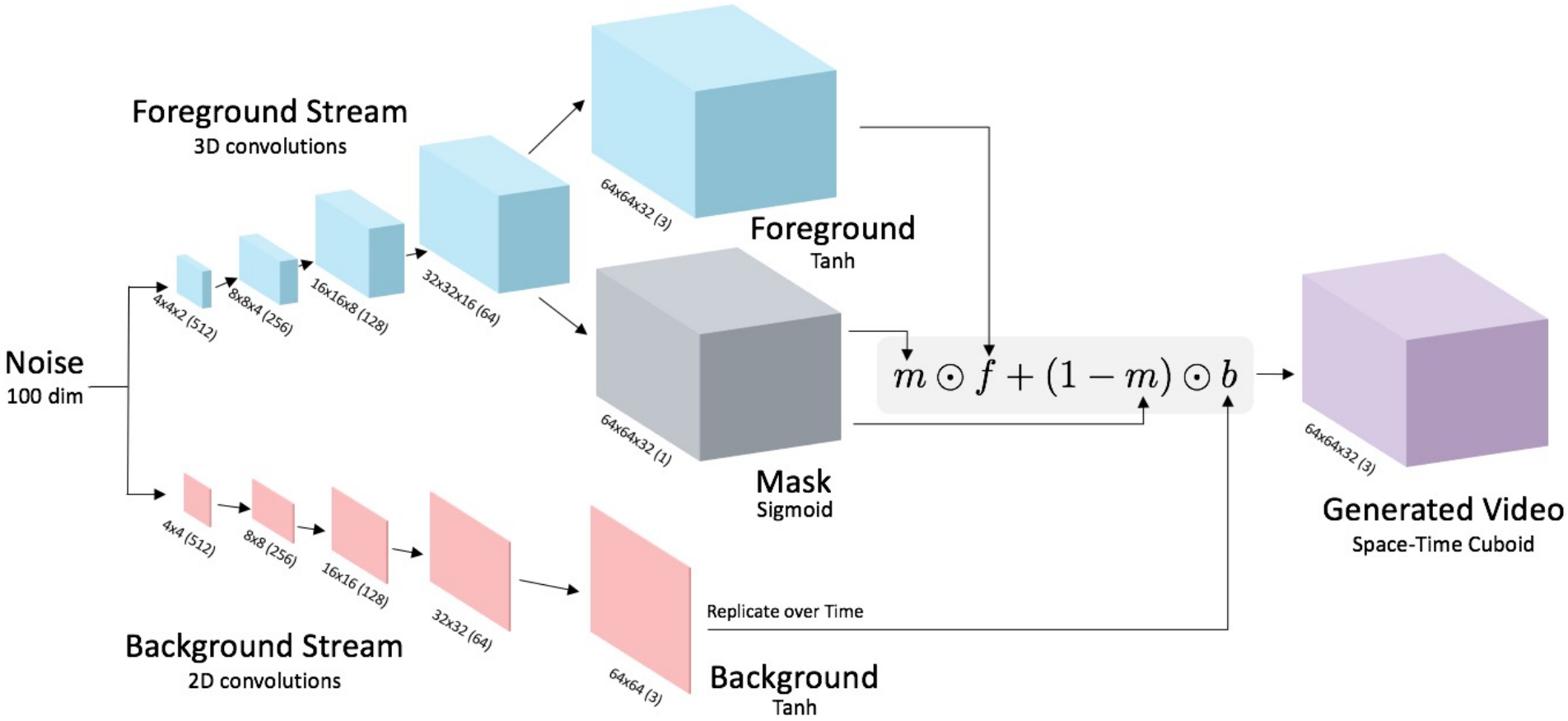
Output Image

Guidance Gradient



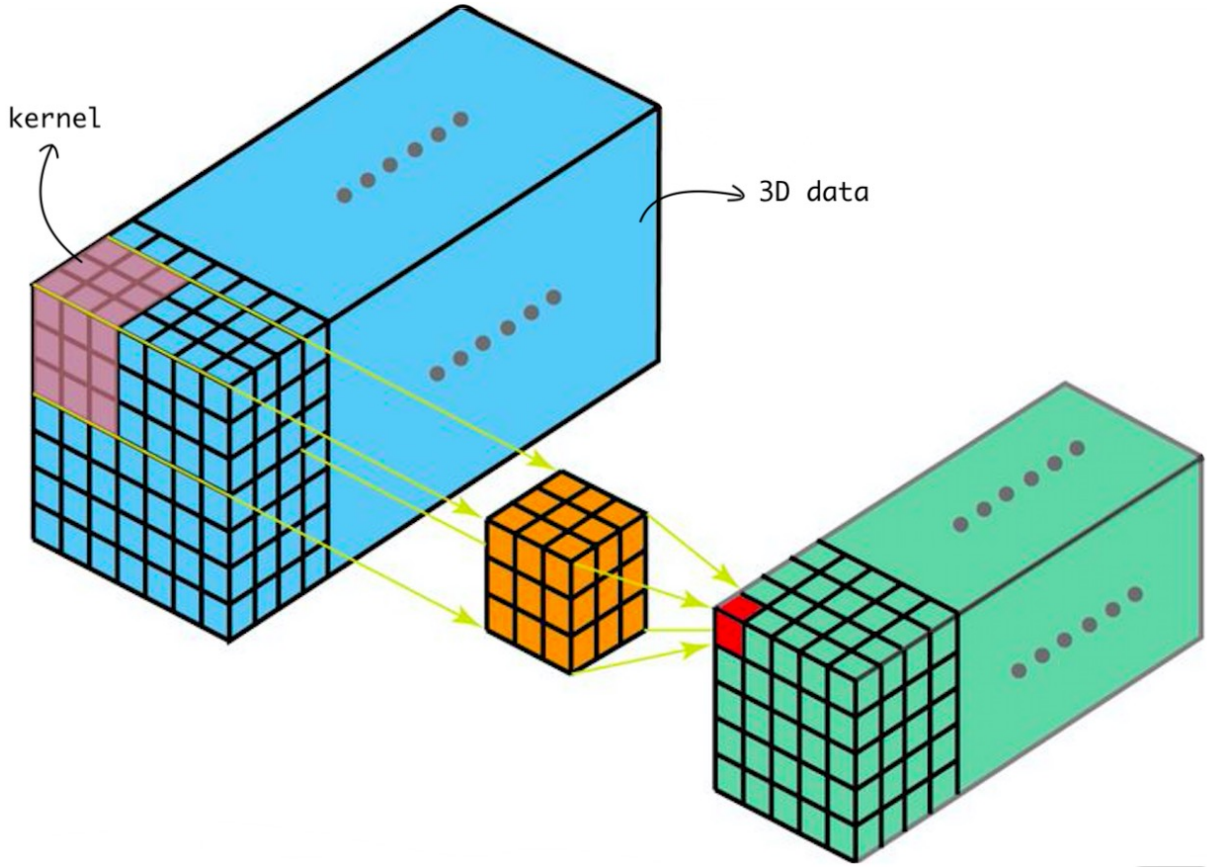
H. Wang, R. Raskar, and N. Ahuja, “Seamless video editing,” in ICPR 2004

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

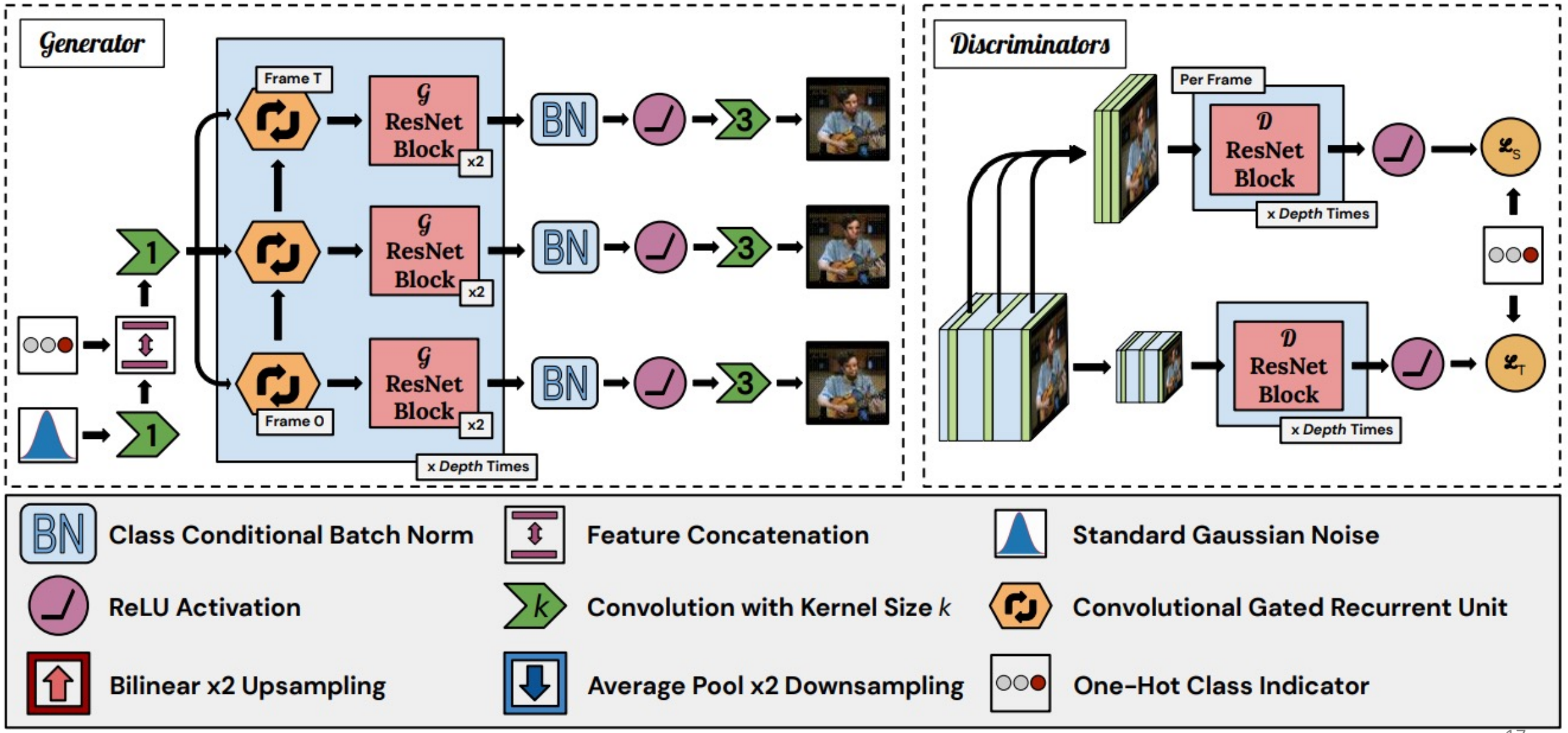
Vondrick et al., NeurIPS 2016

Case Study: DVDGANs

BigGAN-based generator + Spatial Discriminator + Temporal Discriminator

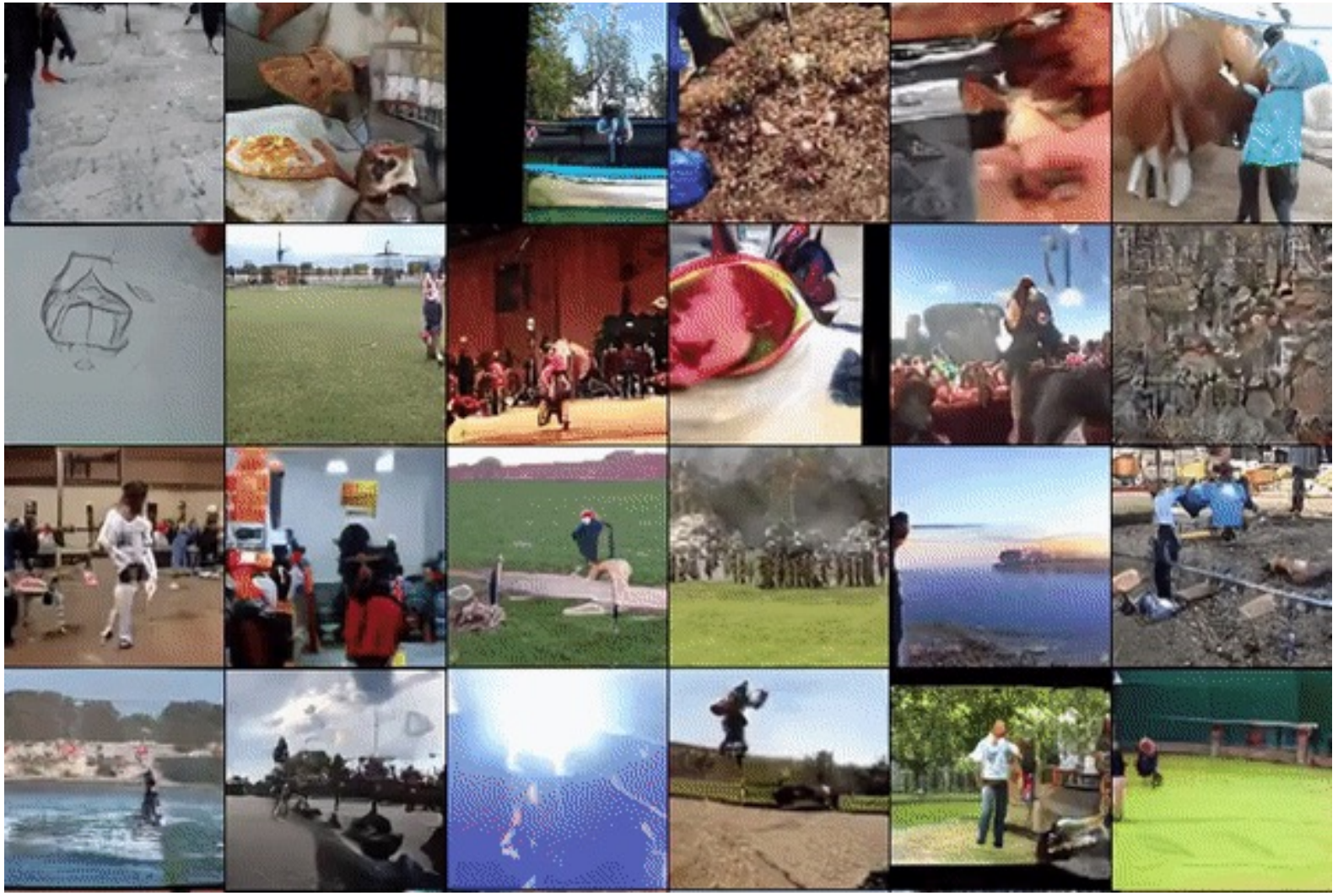
Clark et all, arXiv 2019

Case Study: DVDGANs



Clark et all, arXiv 2019

Case Study: DVDGANs



Clark et all, arXiv 2019

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 1st frame, generate 2nd frame based on 1st one, …
  - Using optical flow (optional): warp 1st frame to 2nd 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

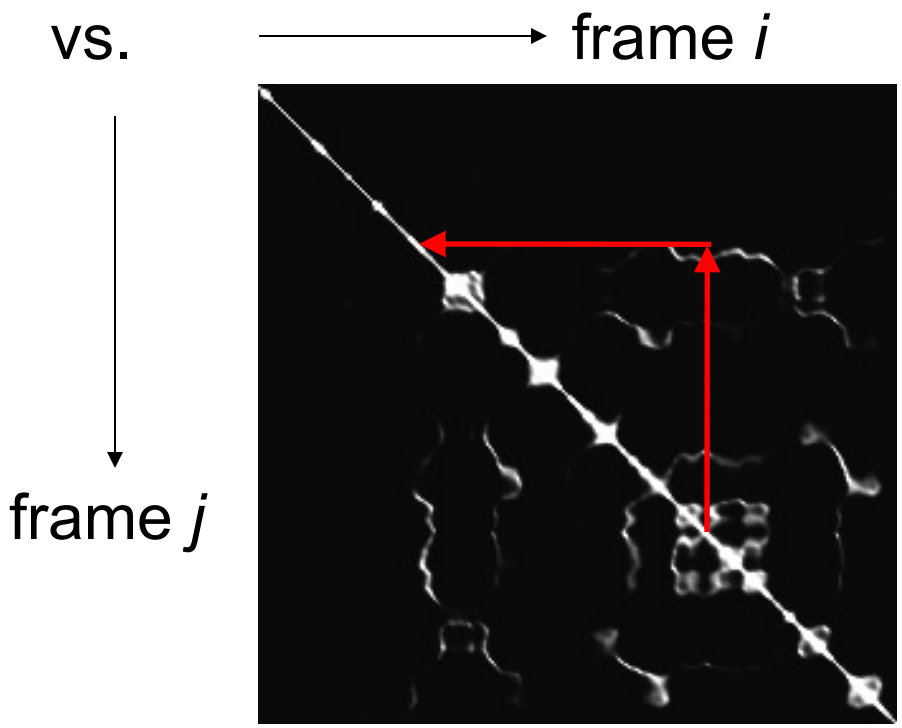
[Schödl et al., SIGGRAPH 2000]

Case Study: Video Textures

[Schödl et al., SIGGRAPH 2000]

Finding good transitions

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



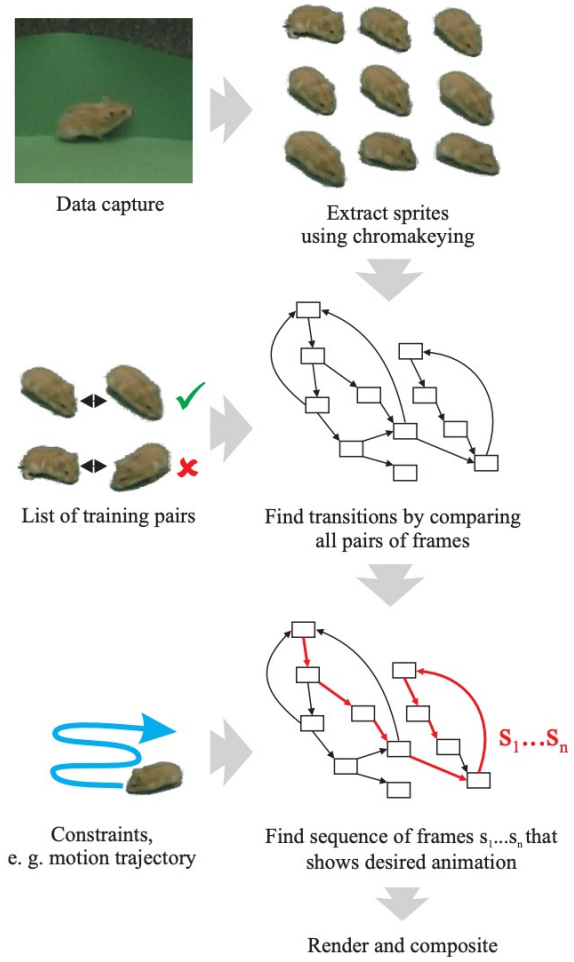
Similar frames make good transitions

[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

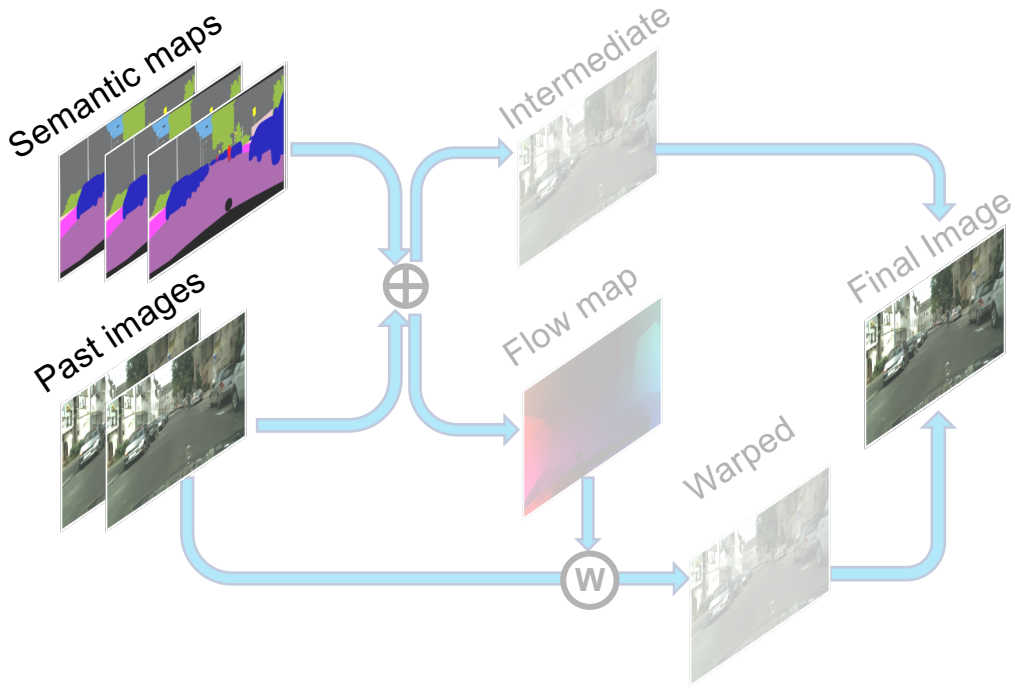
vid2vid

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

Vid2vid [Wang et al., NeurIPS 2018]

vid2vid

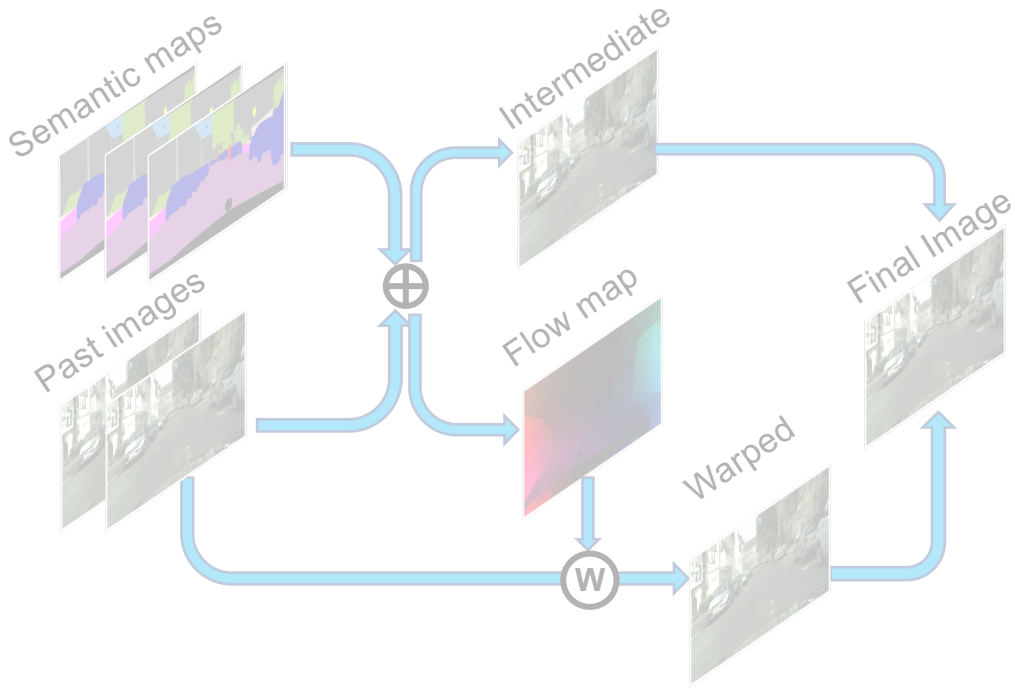
Sequential Generator



Vid2vid [Wang et al., NeurIPS 2018]

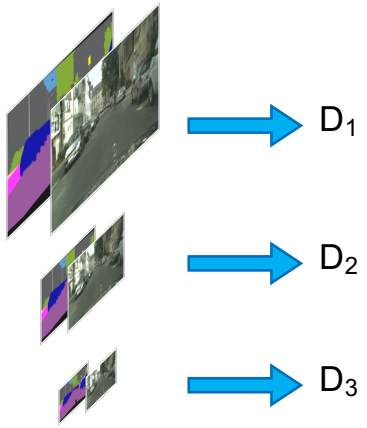
vid2vid

Sequential Generator

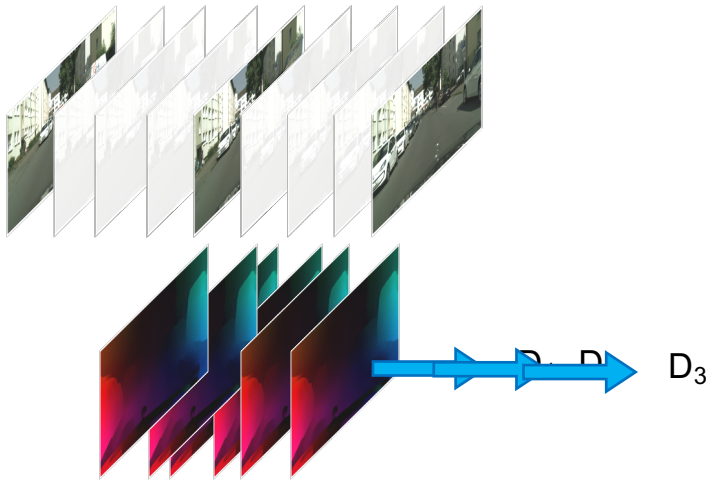


Multi-scale Discriminators

Image Discriminator



Video Discriminator

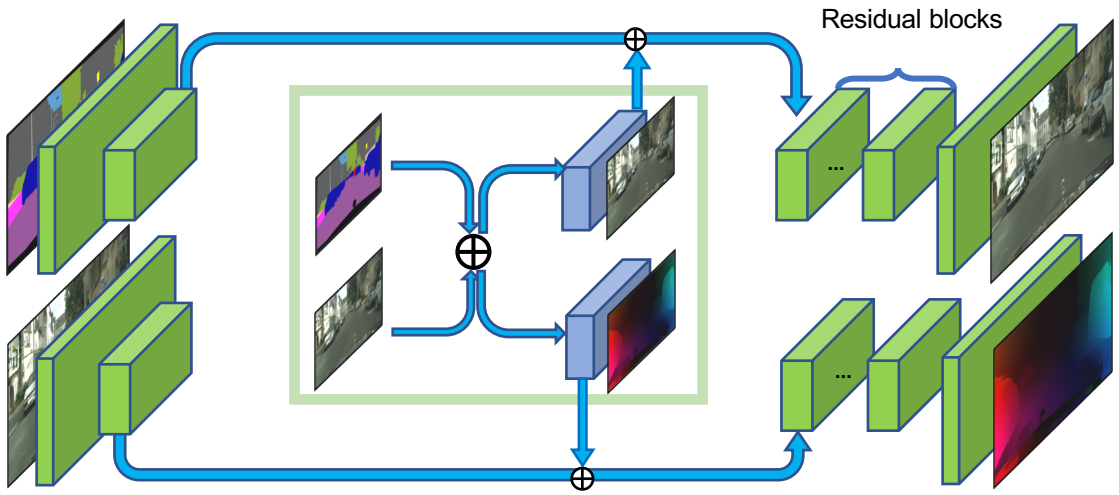


Vid2vid [Wang et al., NeurIPS 2018]

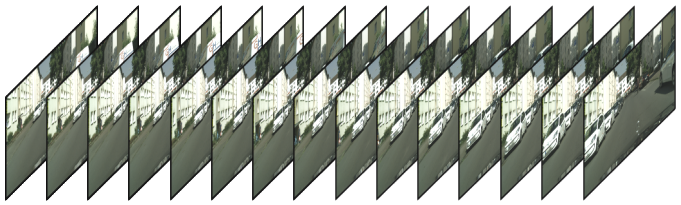
vid2vid

Spatio-temporally Progressive Training

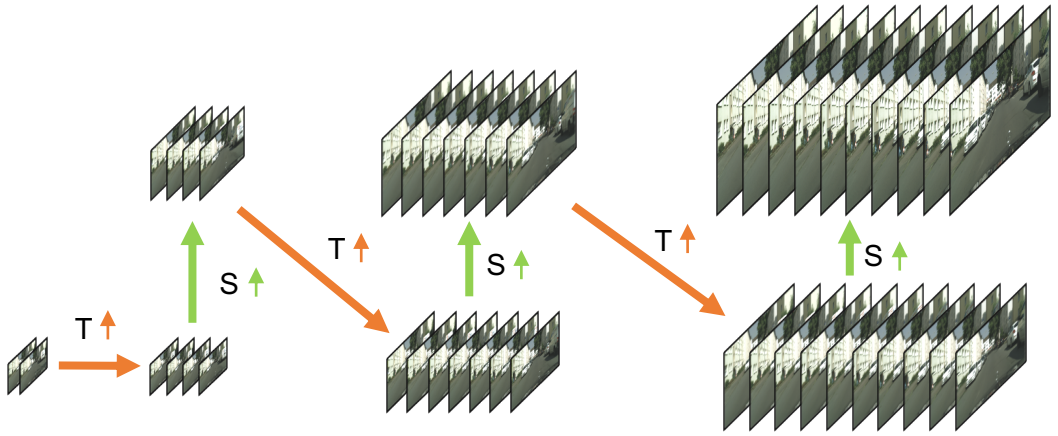
Spatially progressive



Temporally progressive



Alternating training



Vid2vid [Wang et al., NeurIPS 2018]

vid2vid Results

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

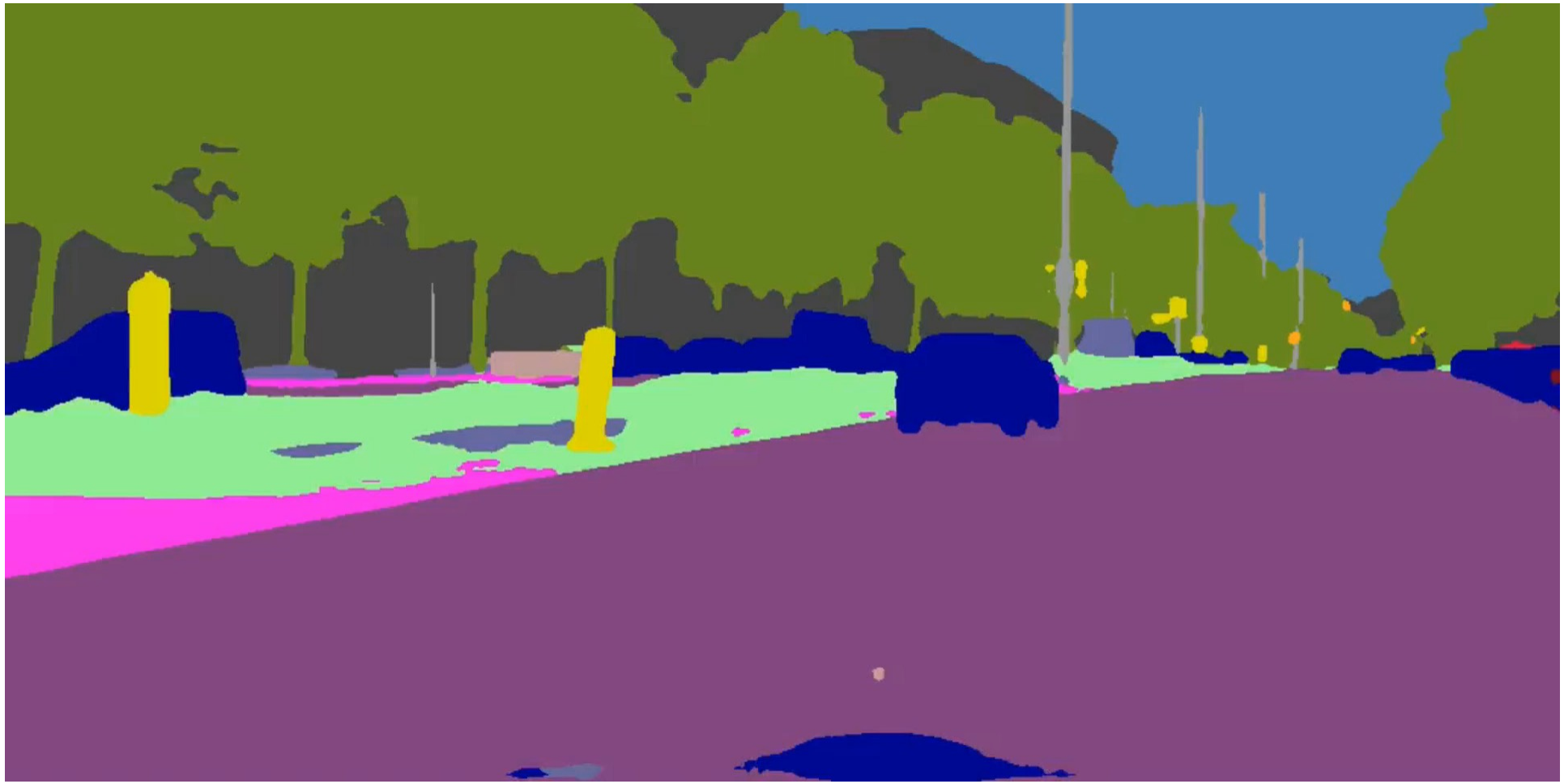
Vid2vid [Wang et al., NeurIPS 2018]

vid2vid Results

- Semantic → Street view scenes

Vid2vid [Wang et al., NeurIPS 2018]

Street View: Cityscapes

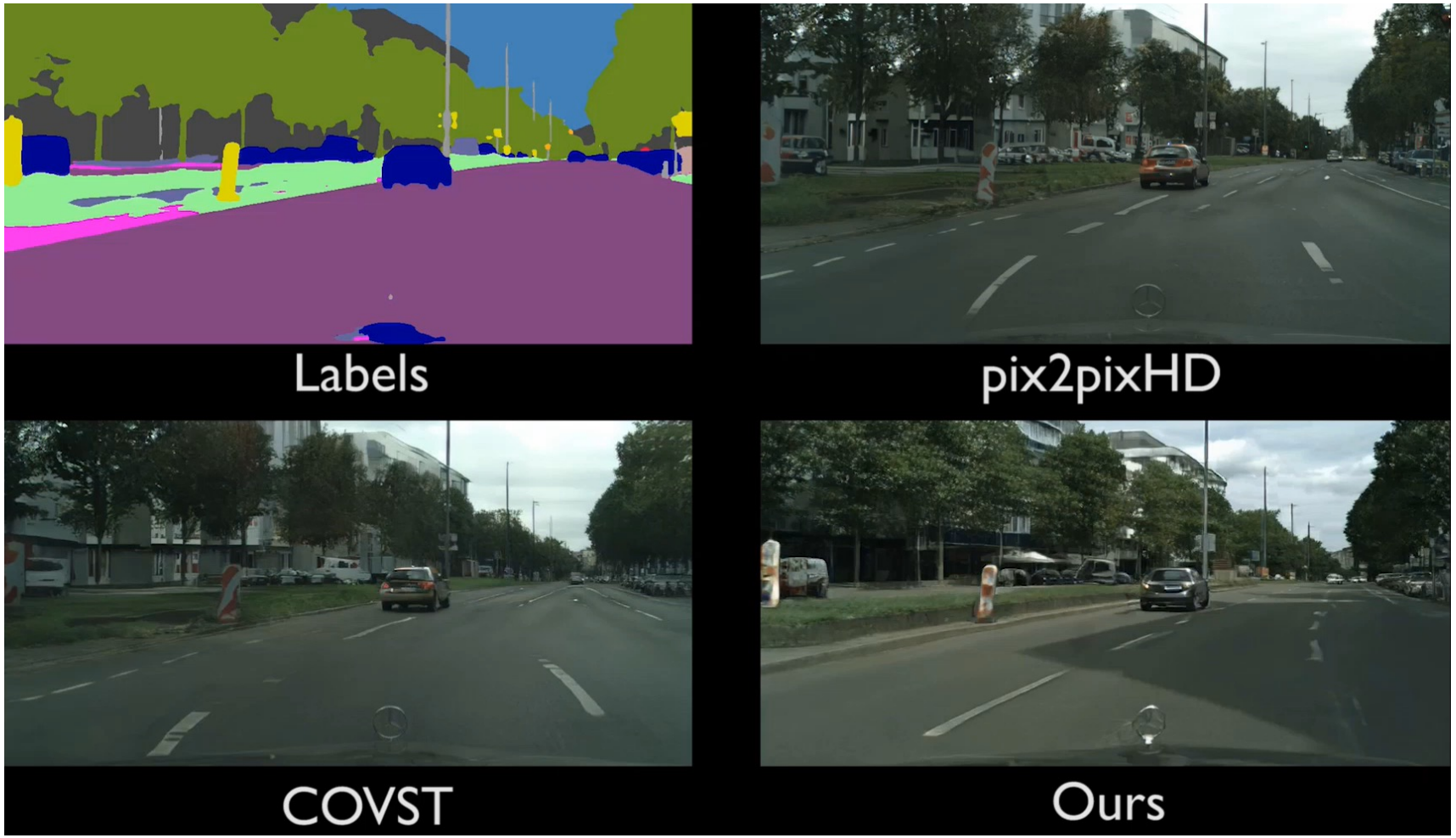


Vid2vid [Wang et al., NeurIPS 2018]

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

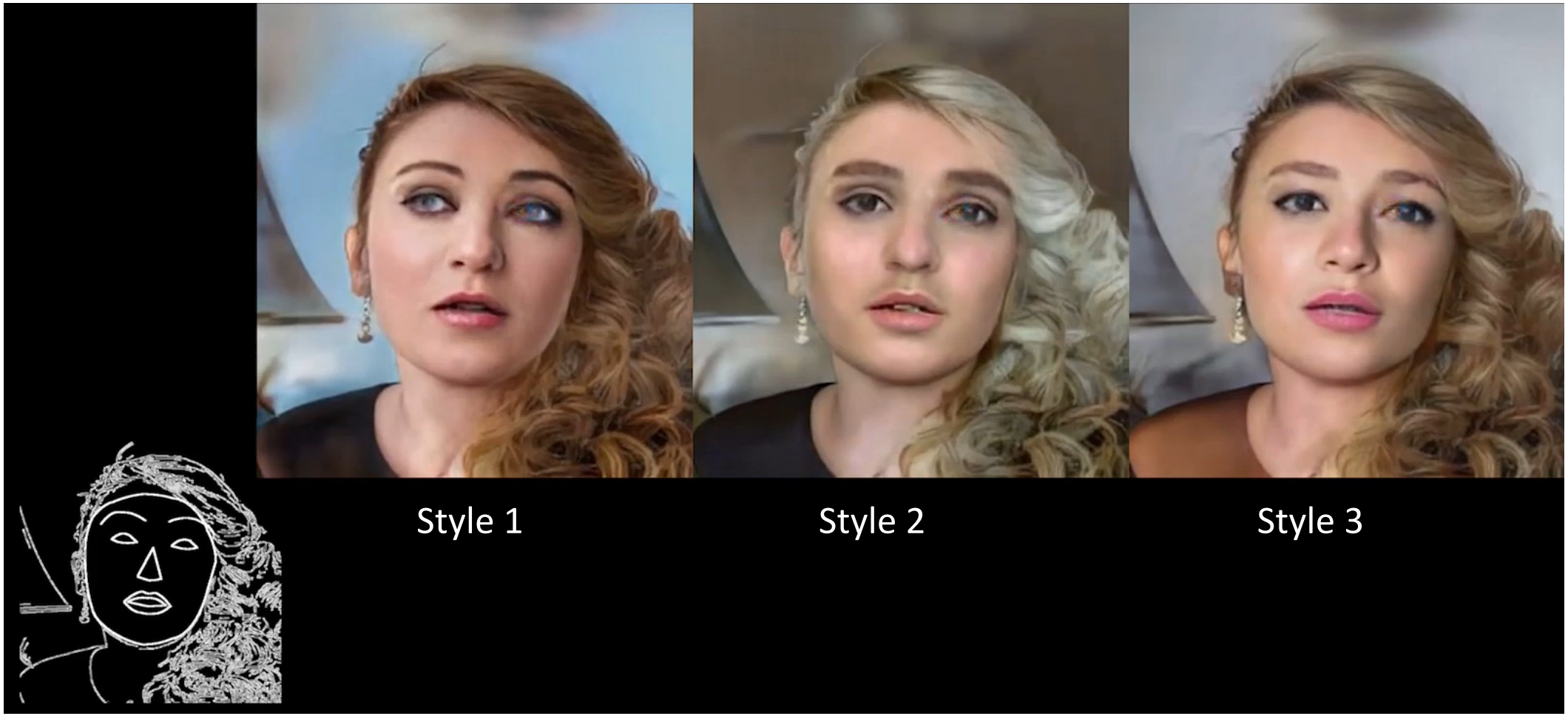
- Edges → Human faces

Face Swapping (face → edge → face)

input edges output

Vid2vid [Wang et al., NeurIPS 2018]

Multi-modal Edge → Face



Results

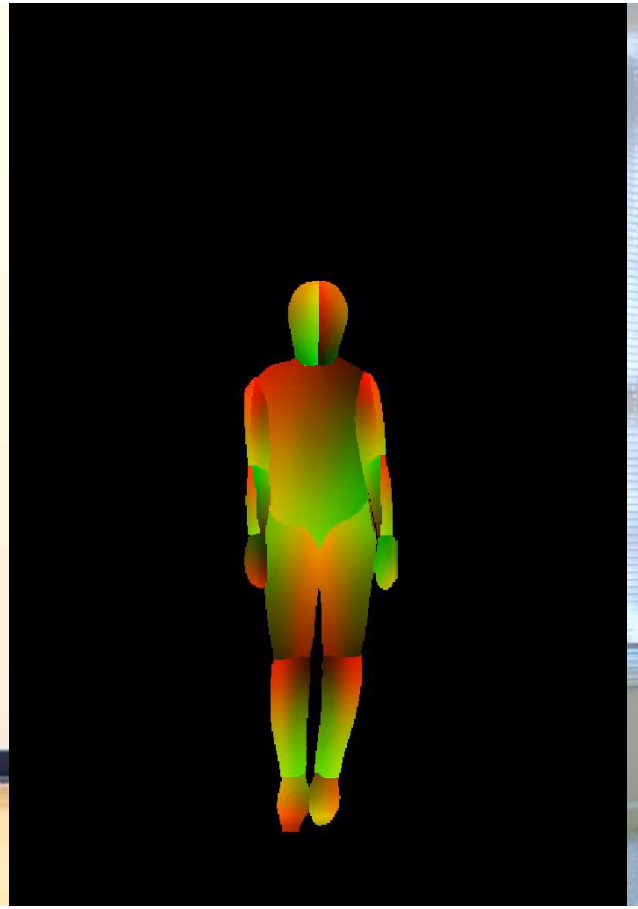
- Poses → Human bodies

Motion Transfer (body → pose → body)

input poses output

Motion Transfer (body → pose → body)

input



poses

output

Motion Transfer (body → pose → body)

input poses output

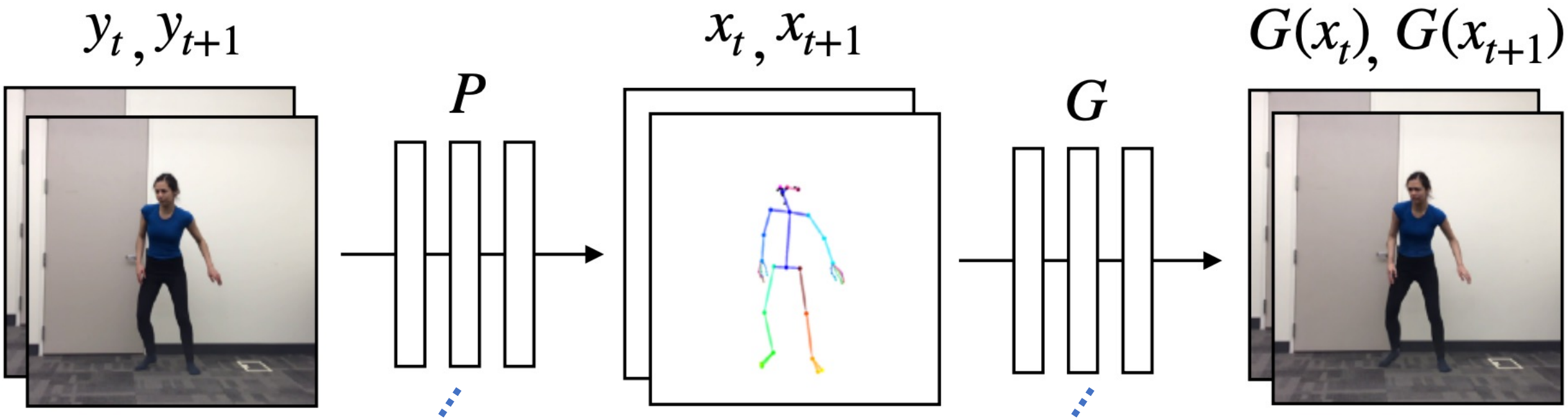
Motion Transfer (body → pose → body)

input poses output

More Dancing …

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

Pose-guided Synthesis

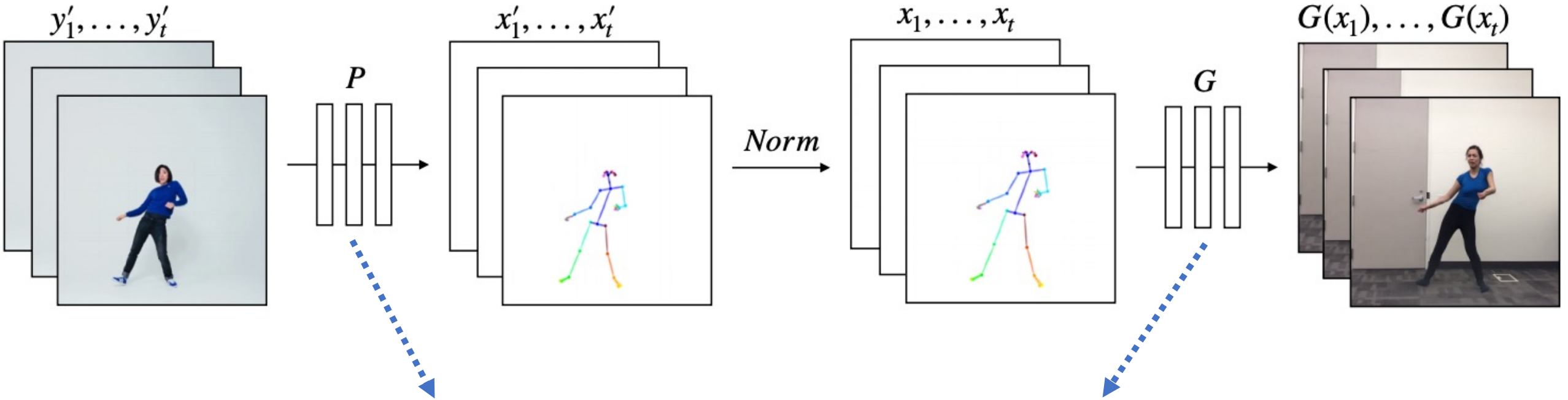


Pose network

Generator network

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

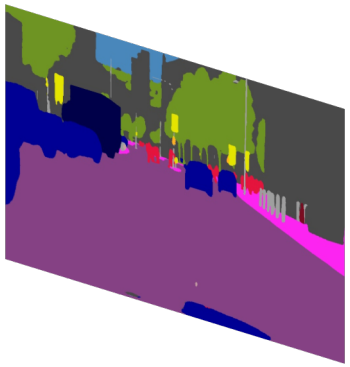
Data

Vid2vid [Wang et al., NeurIPS 2018]

vid2vid Extensions: Interactive Graphics

User control

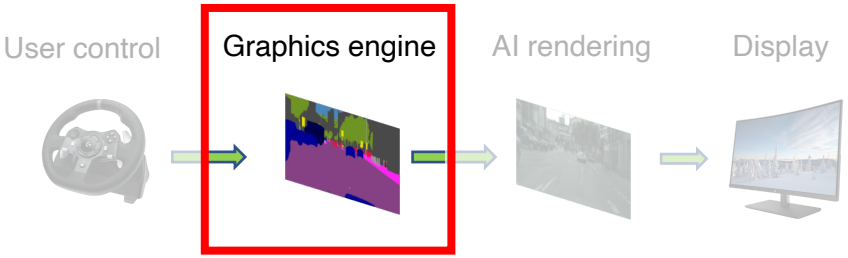
Graphics engine



AI rendering

Display

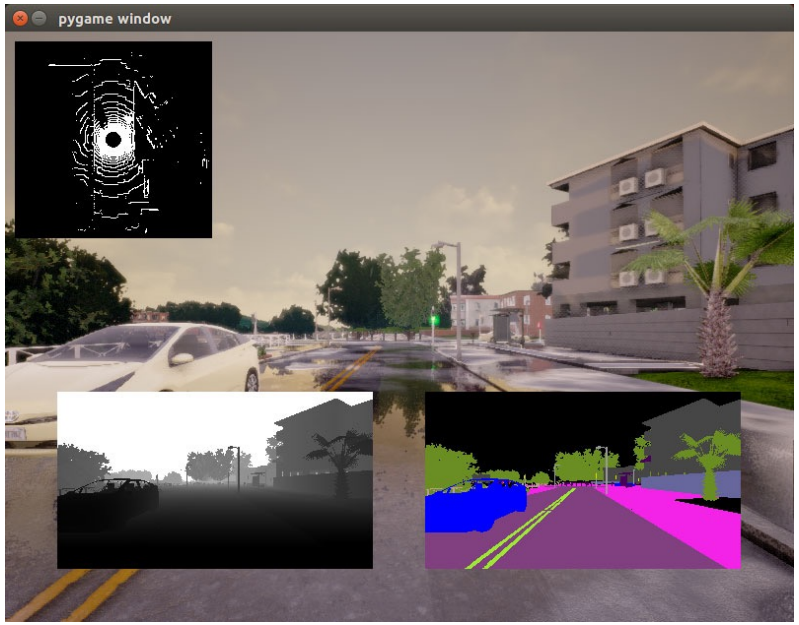
Graphics Engine: CARLA



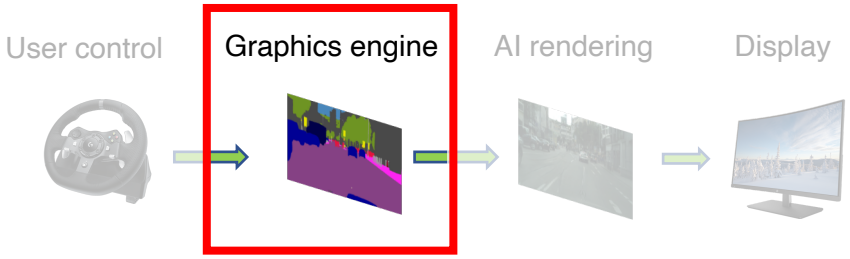
DATA

PROTOBUF messages

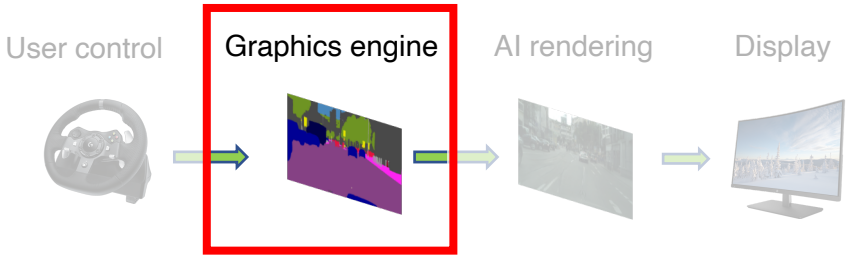
CONTROL

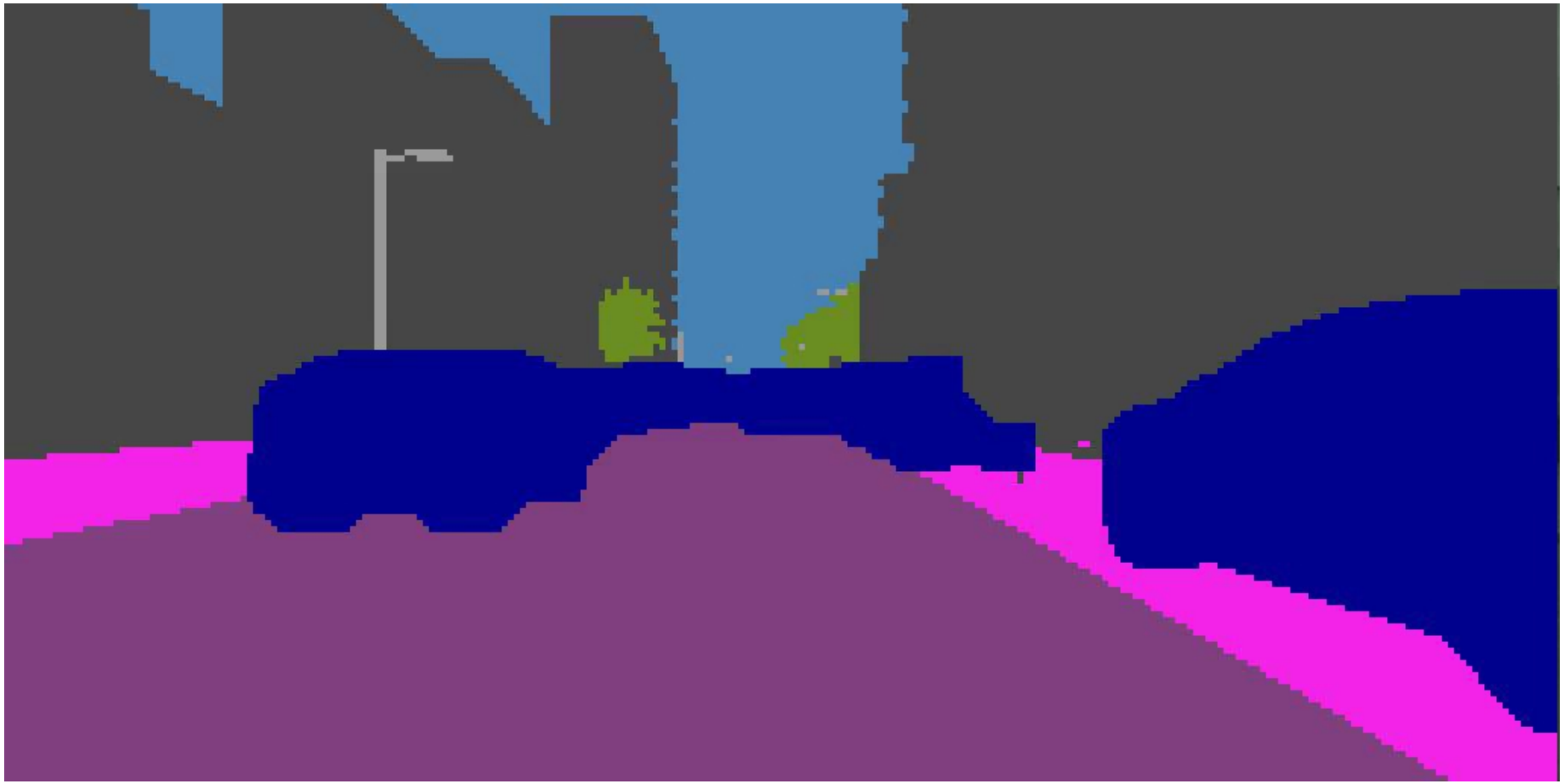


Original CARLA Sequence

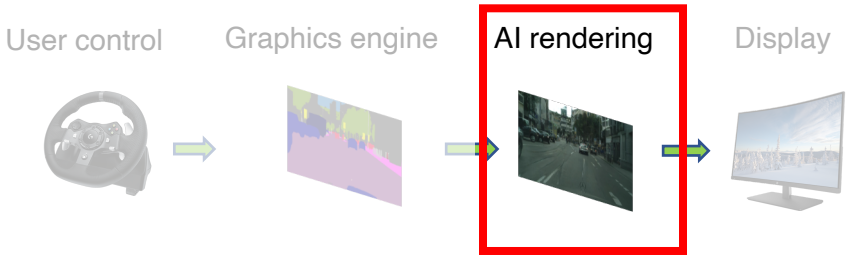


CARLA Semantic Maps

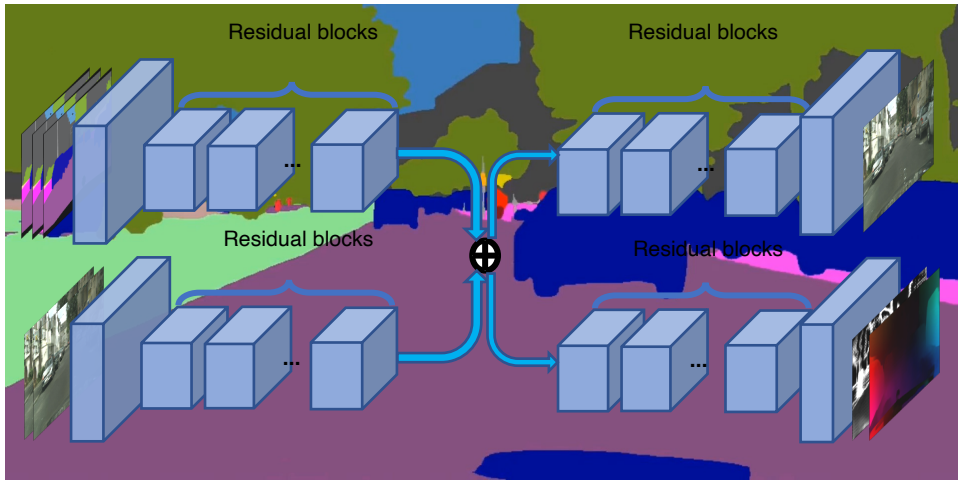


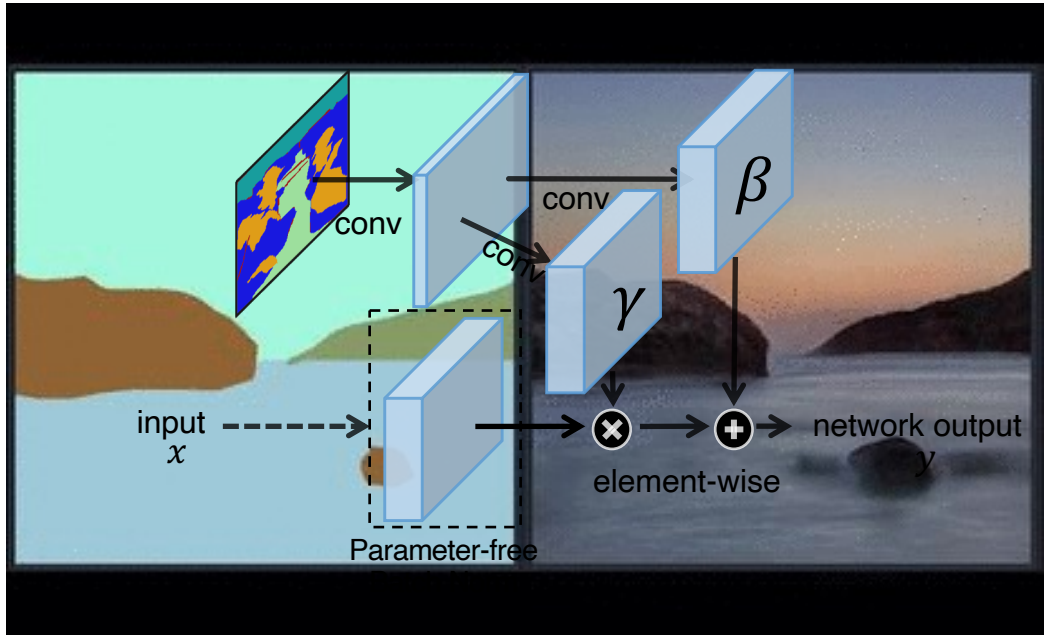


Methodology

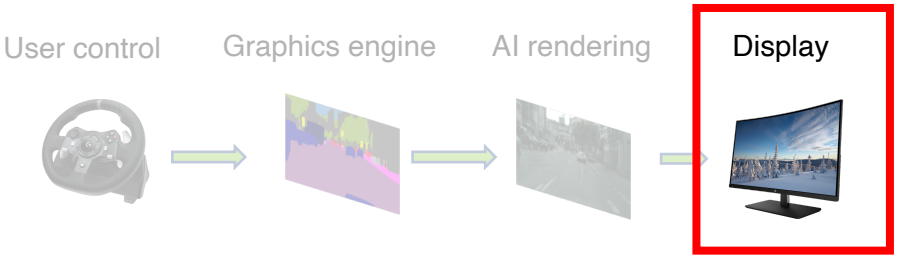


- Combine vid2vid with SPADE





Demo Result



Driving Game

vid2game by FAIR

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

Adaptive Video Translation

Disadvantages of vid2vid

- Separate models for each dataset

model 1

model 2

model 3

- Generalizing to new persons requires

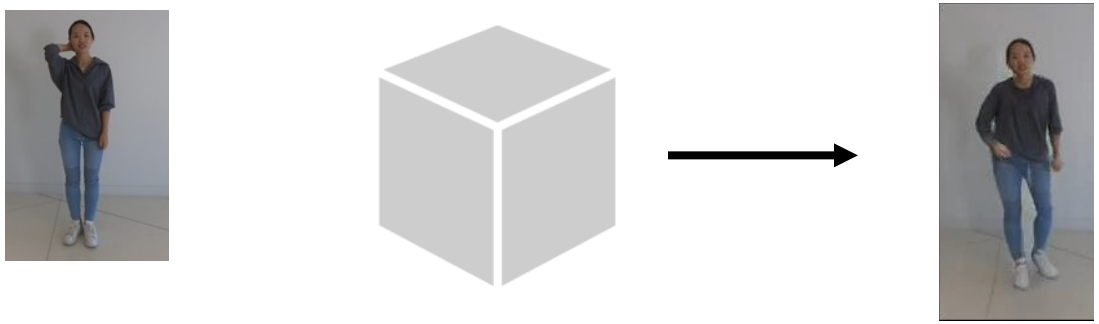
Collecting new data

Training

Wouldn’t it be great if…

- One model for all

- Dynamically determine the style at run time
  - based on an *exemplar image*



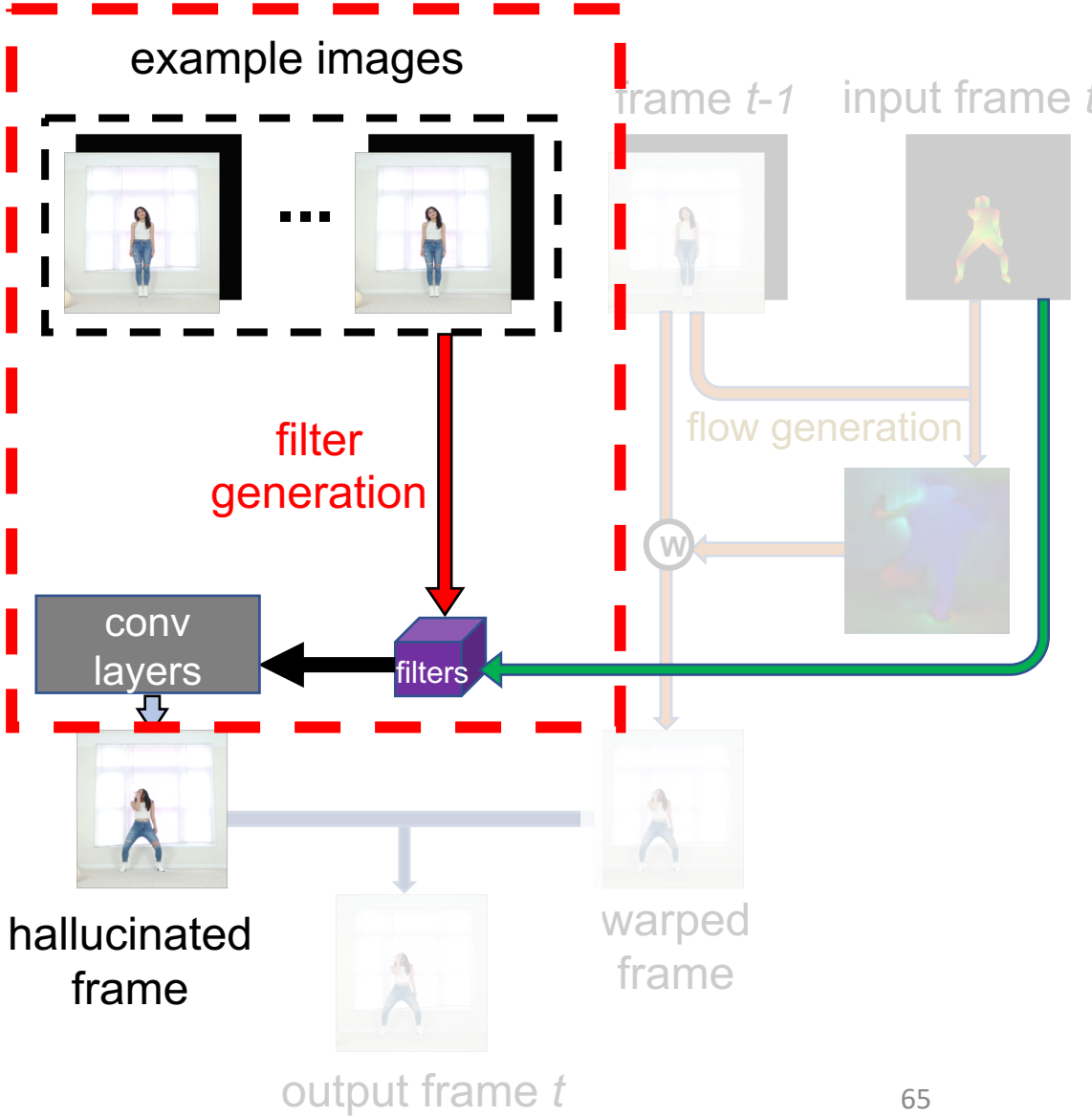
Adaptive Video-to-Video Translation

T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, B. Catanzaro, “Few-shot Adaptive Video-to-Video Synthesis,” NeurIPS 2019.

Adaptive vid2vid: overflow

- Original vid2vid
  - Output frame =
    Hallucinated frame + Warped frame

- Adaptive vid2vid
  - Hallucinated frames
    - generated based on example images
  - Using a filter generation scheme

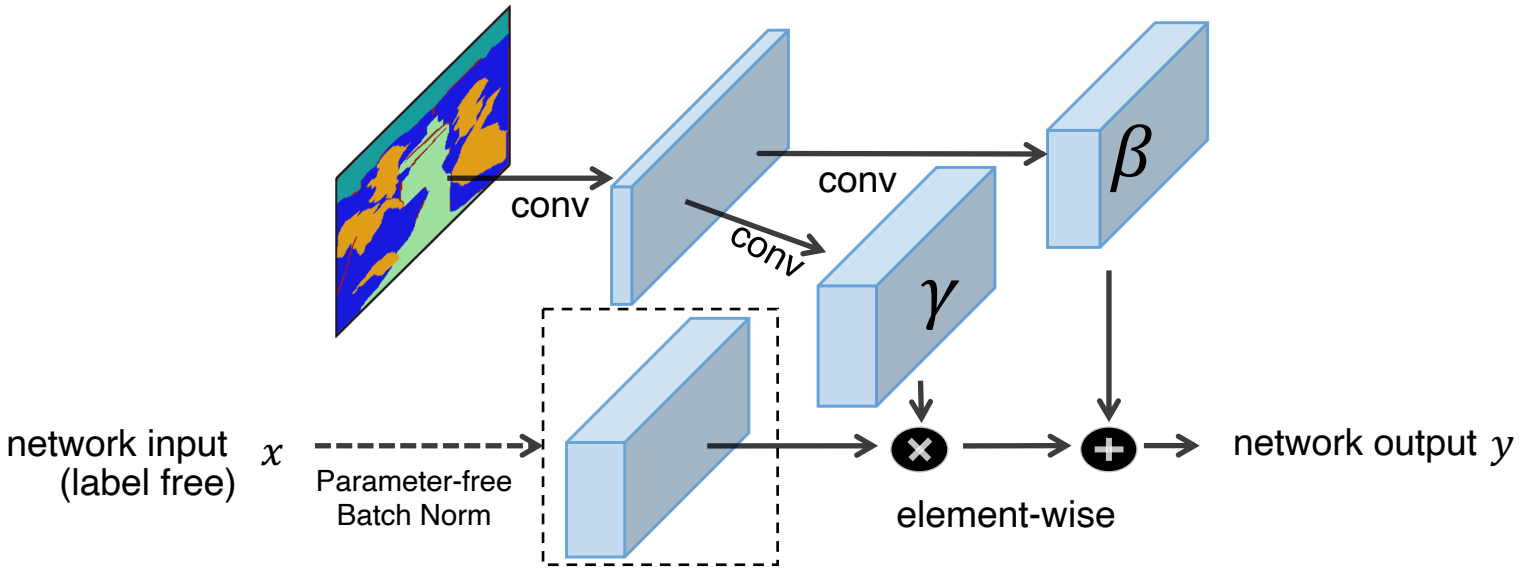


Adaptive vid2vid

- Based on SPADE (GauGAN)
  - ~~Prior work: input semantics → encoder-decoder → output image~~
  - Instead: input semantics

    → ***spatially-varying*** normalization maps

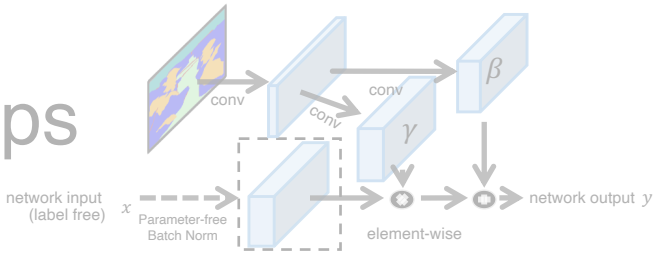
    → used in every BatchNorm



$$y = \frac{x - \mu}{\sigma} \cdot \gamma + \beta$$

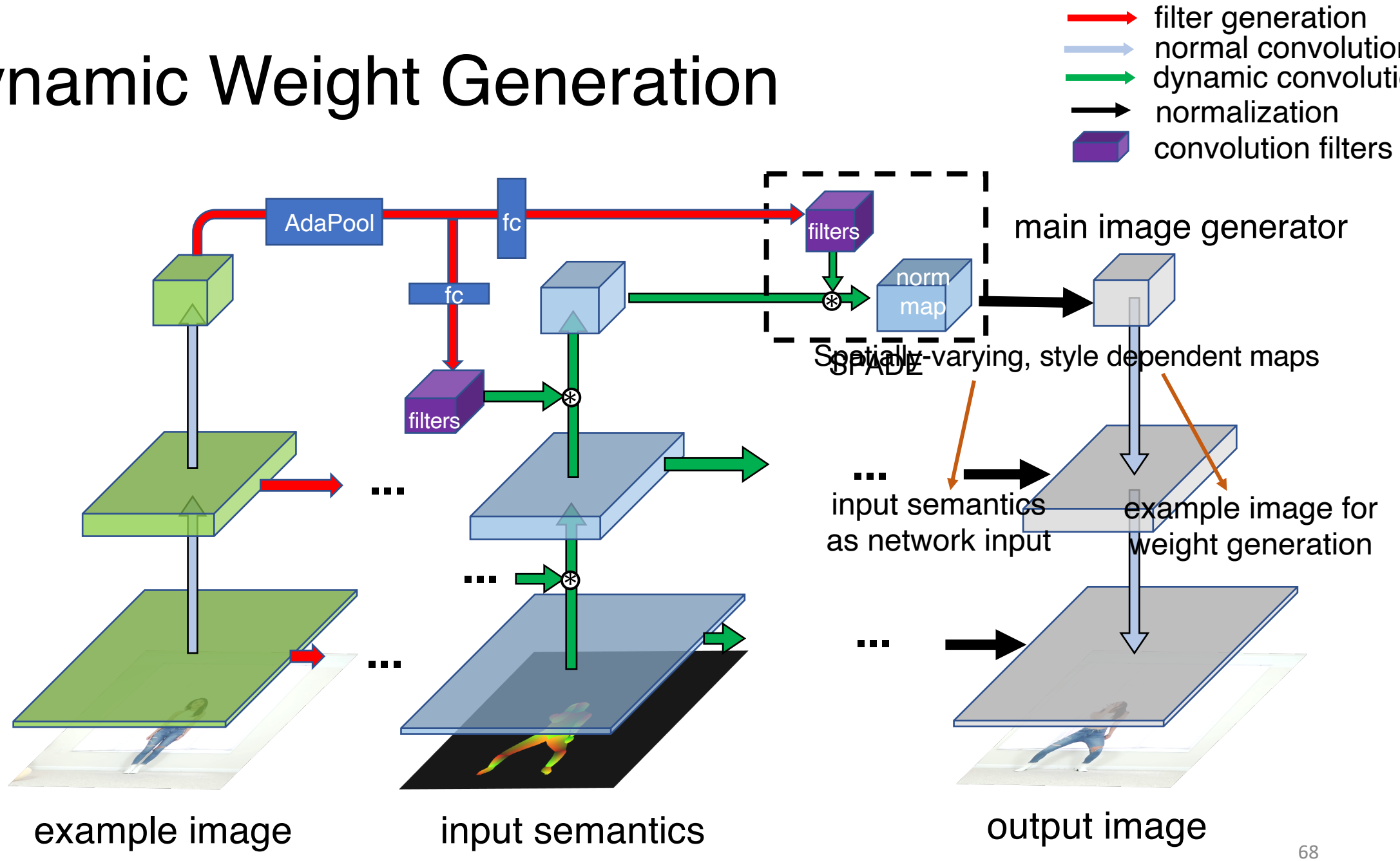
Adaptive vid2vid

- Based on SPADE (GauGAN)
  - Prior work: input semantics → encoder-decoder → output image
  - Instead: input semantics
    → ***spatially-varying*** normalization maps
    → used in every BatchNorm



- Given an additional exemplar image
  - Dynamically configure the ***network weights*** in SPADE
  - Generate ***spatially-varying, style-dependent*** normalization maps
    - Spatial info ← input semantics
    - Style info ← exemplar images

Dynamic Weight Generation



Adaptive vid2vid: Training

- From a video
  - Randomly sample a clip
  - Randomly sample another reference frame(s)

- Make the network generate the clip
  - Based on the reference frame

Adaptive vid2vid: Testing

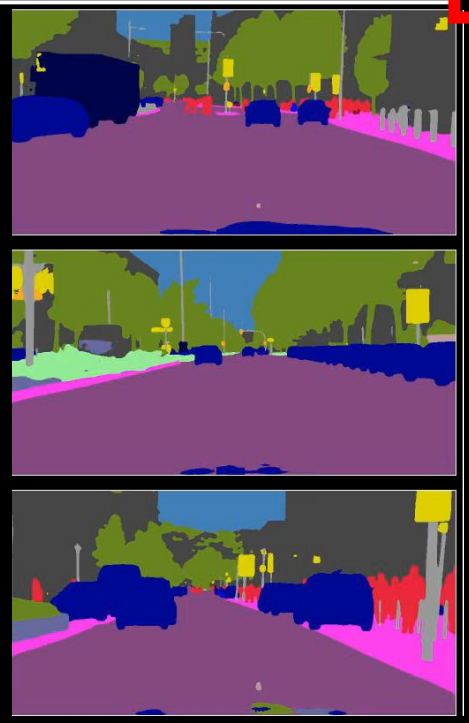
- Given an example image
- Finetune on the example image
  - Network output should be the same as the example
  - Only finetune for a few iterations

- For faces: normalize keypoints
  - To the same as example image
  - To better preserve identity

Results

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

Street View Scenes



Input segmentations

Edges → Faces

Example images

Edges → Faces

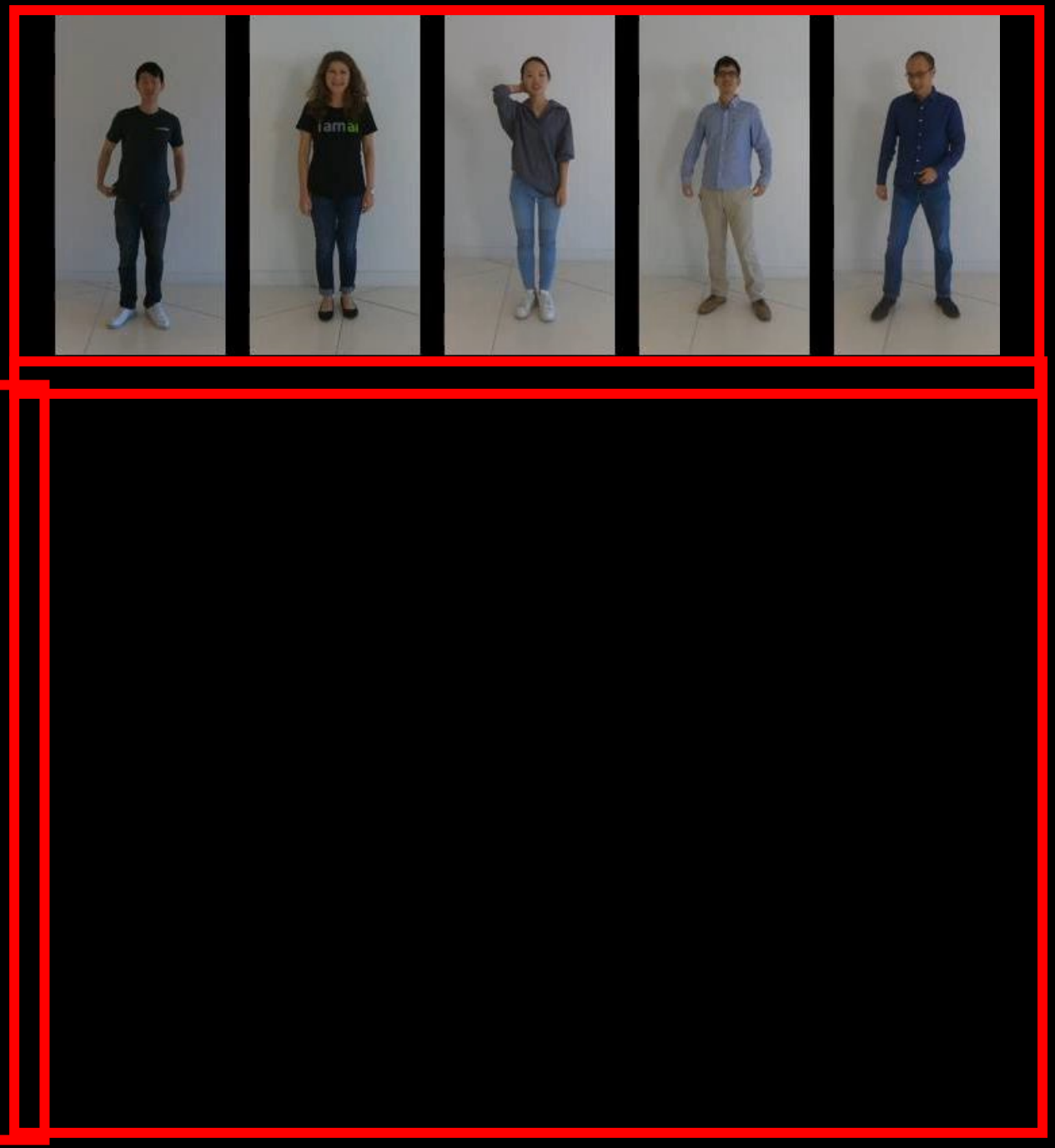
Example image

Input videos

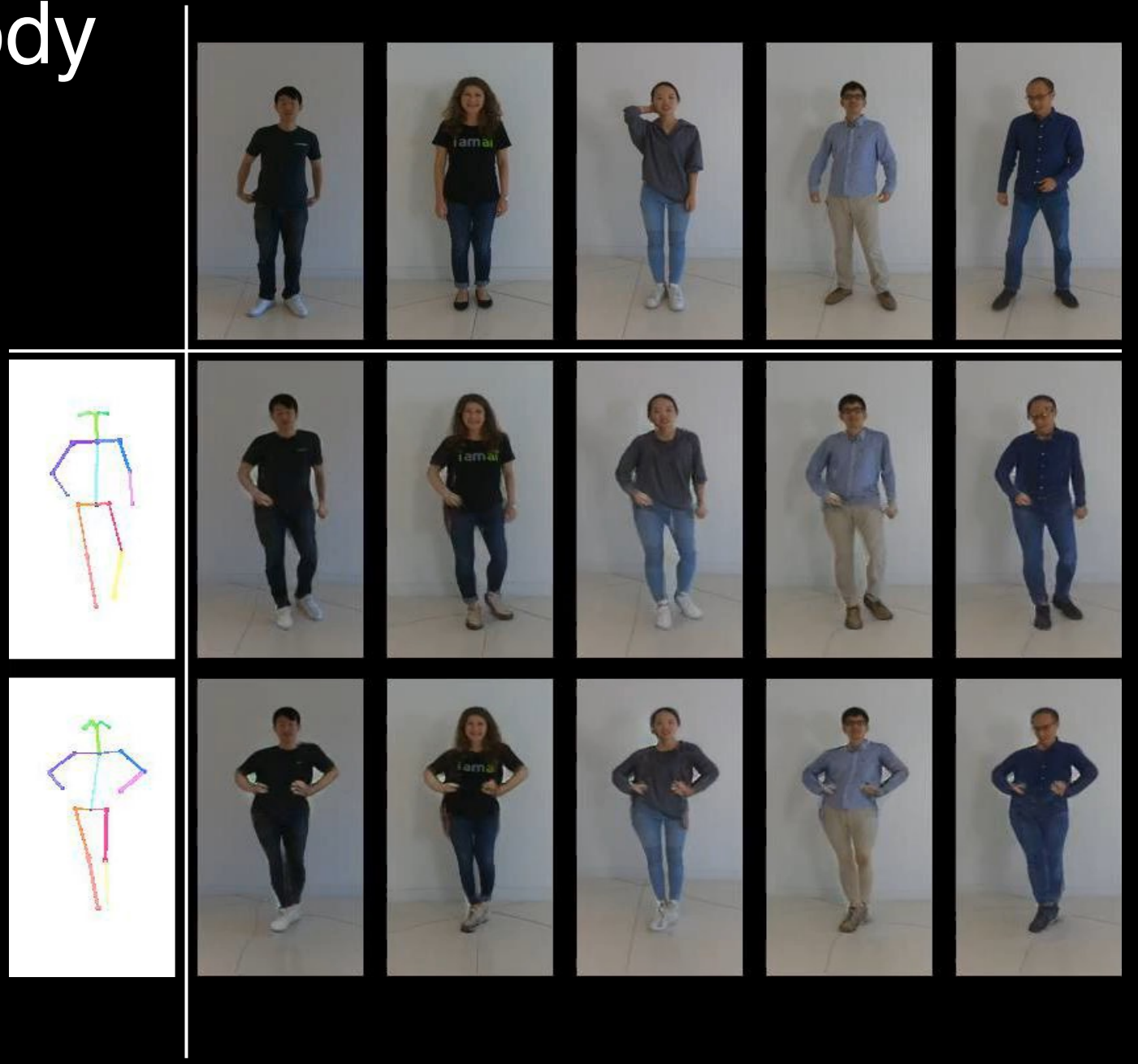
Extracted edges

Synthesized result

Poses → Body



Poses → Body



Poses → Body

Poses → Body

Example image

Input videos

Poses

Synthesized

Face Video Compression

[Wang et al. CVPR 2021]

Paired vs. Unpaired

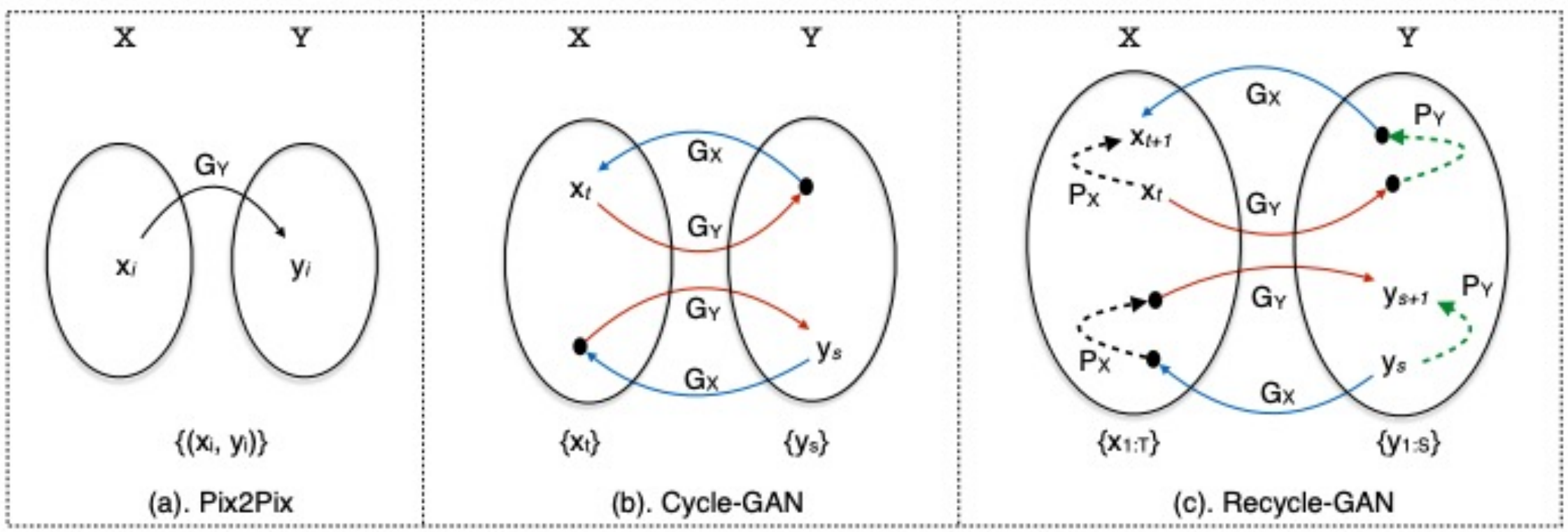
Unpaired Video Learning with RecycleGAN

John Oliver to Stephen Colbert

Flower to Flower

[Bansal et al., ECCV 2018]

Unpaired Video Learning with RecycleGAN



Prediction Loss $$L_\tau(P_X) = \sum_t ||x_{t+1} - P_X(x_{1:t})||^2,$$

Recycle Loss $$L_r(G_X, G_Y, P_Y) = \sum_t ||x_{t+1} - G_X(P_Y(G_Y(x_{1:t})))||^2,$$

[Bansal et al., ECCV 2018]

Unpaired Video Learning with RecycleGAN

[Bansal et al., ECCV 2018]

Unpaired Video Learning with RecycleGAN

[Bansal et al., ECCV 2018]

Thank You!

16-726, Spring 2022

https://learning-image-synthesis.github.io/sp22/

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