VIDEO-TO-VIDEO SYNTHESIS

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, Bryan Catanzaro

GENERATIVE ADVERSARIAL NETWORKS

Unconditional GANs

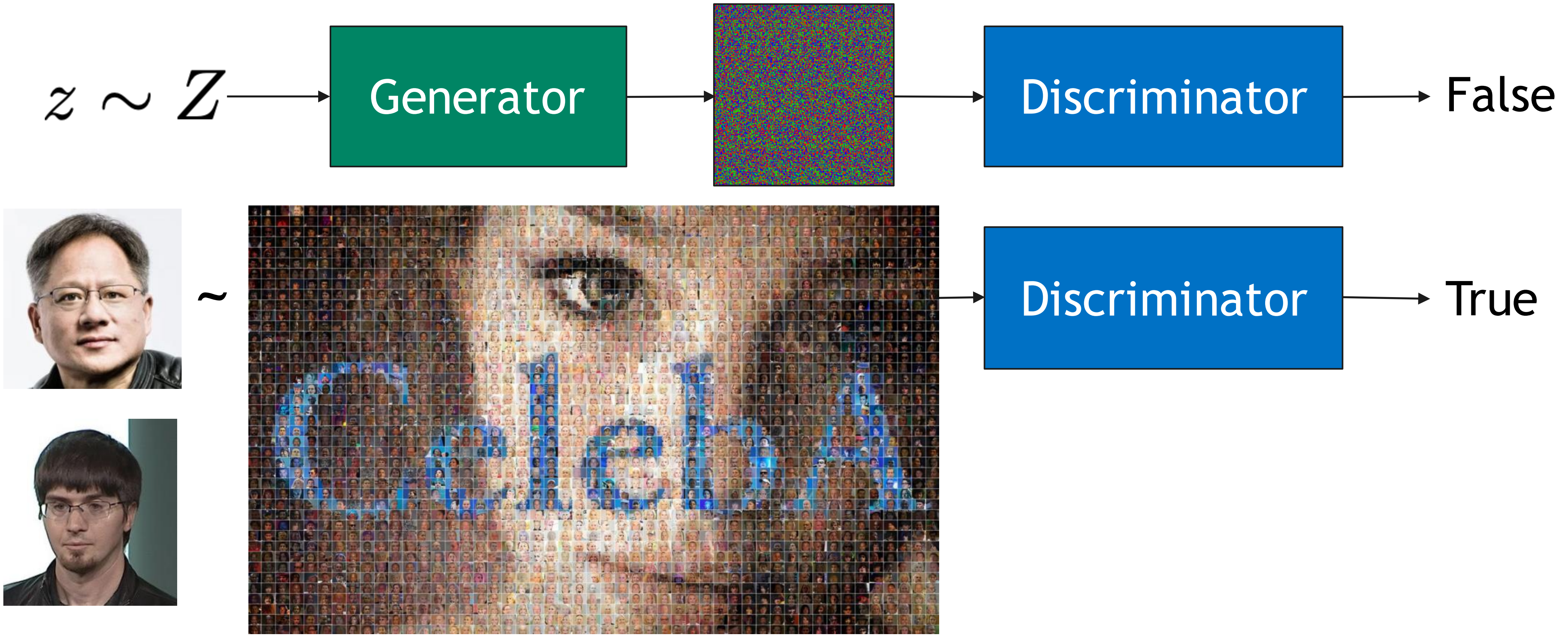


Image credit: Celebrity dataset, Jensen Huang, Founder and CEO of NVIDIA, Ian Goodfellow, Father of GANs.

After training for a while using NVIDIA DGX1 machines

Fun sampling time begin

$z_1, z_2, z_3, \ldots \rightarrow$ Generator $\rightarrow$

Image credit: NVIDIA StyleGAN

CONDITIONAL GANS

Allow user more control on the sampling process

Modeling (training)

$$p_{X|Y}$$

Generated result — Given info (e.g. image, text)

Sampling (testing)

$$z \sim Z, y \sim Y$$

output style — Given info (e.g. image, text)

SKETCH-CONDITIONAL GANS

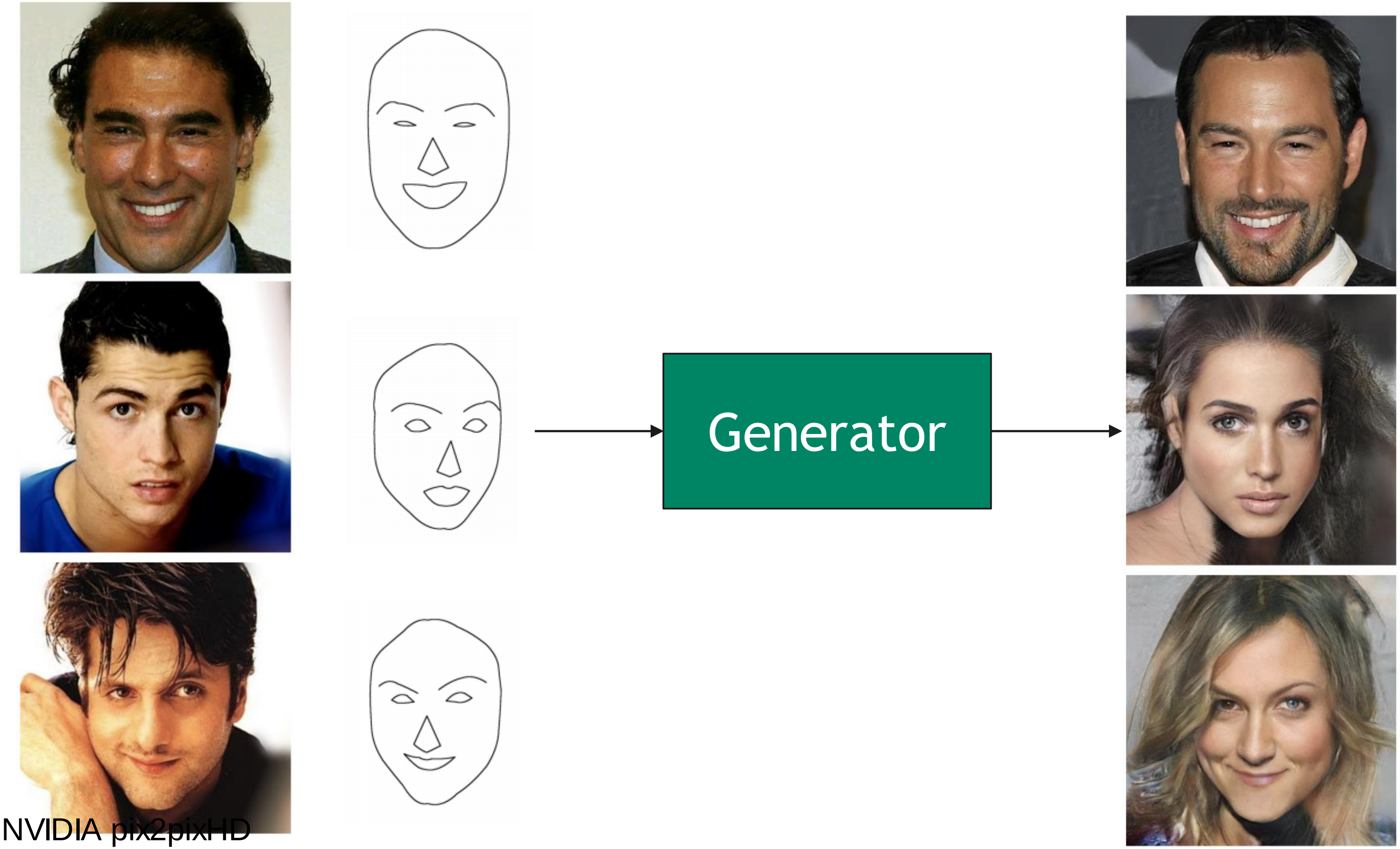


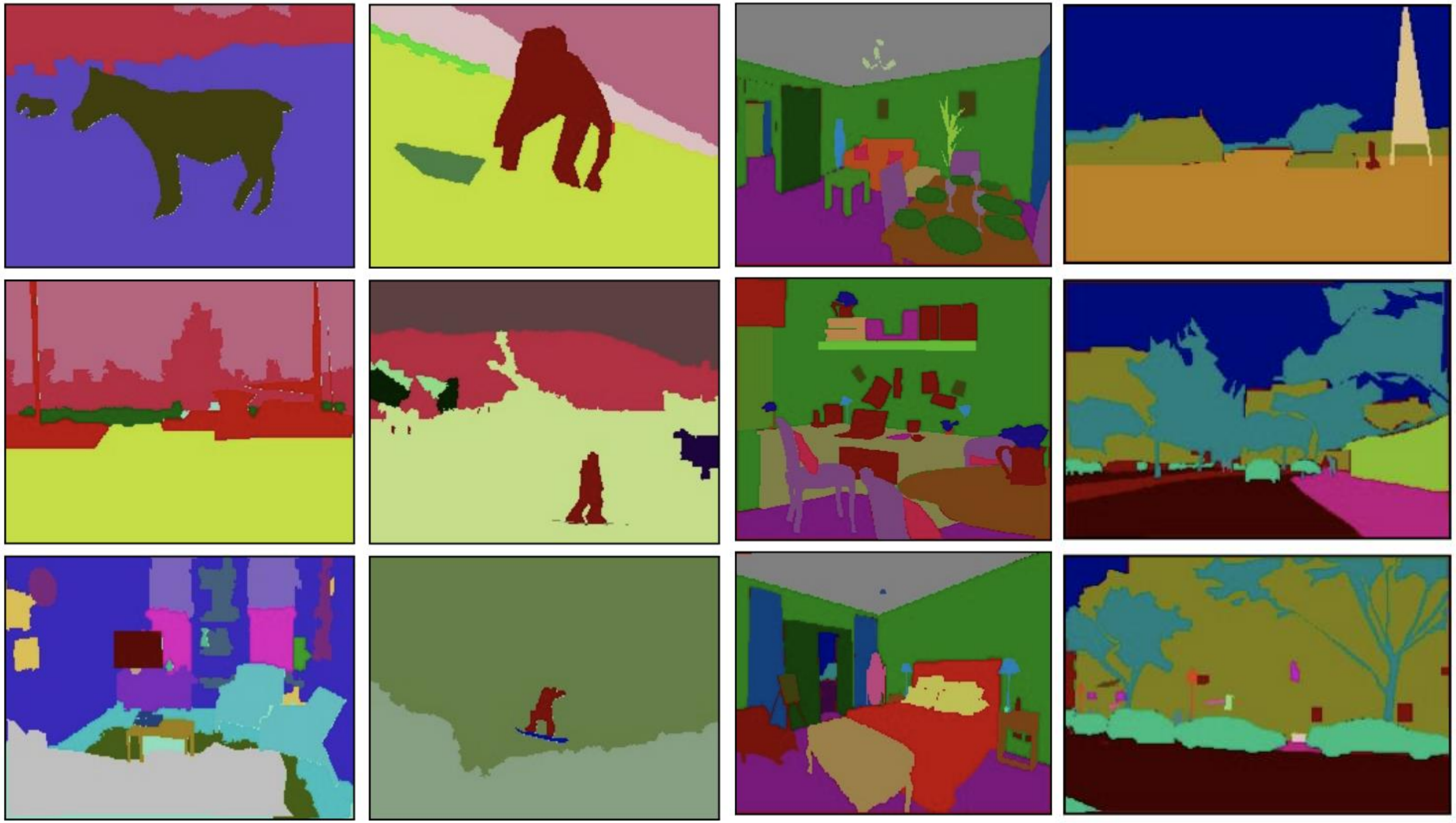
Image credit: NVIDIA pix2pixHD

IMAGE-CONDITIONAL GANS

Image credit: NVIDIA MUNIT

MASK-CONDITIONAL GANS

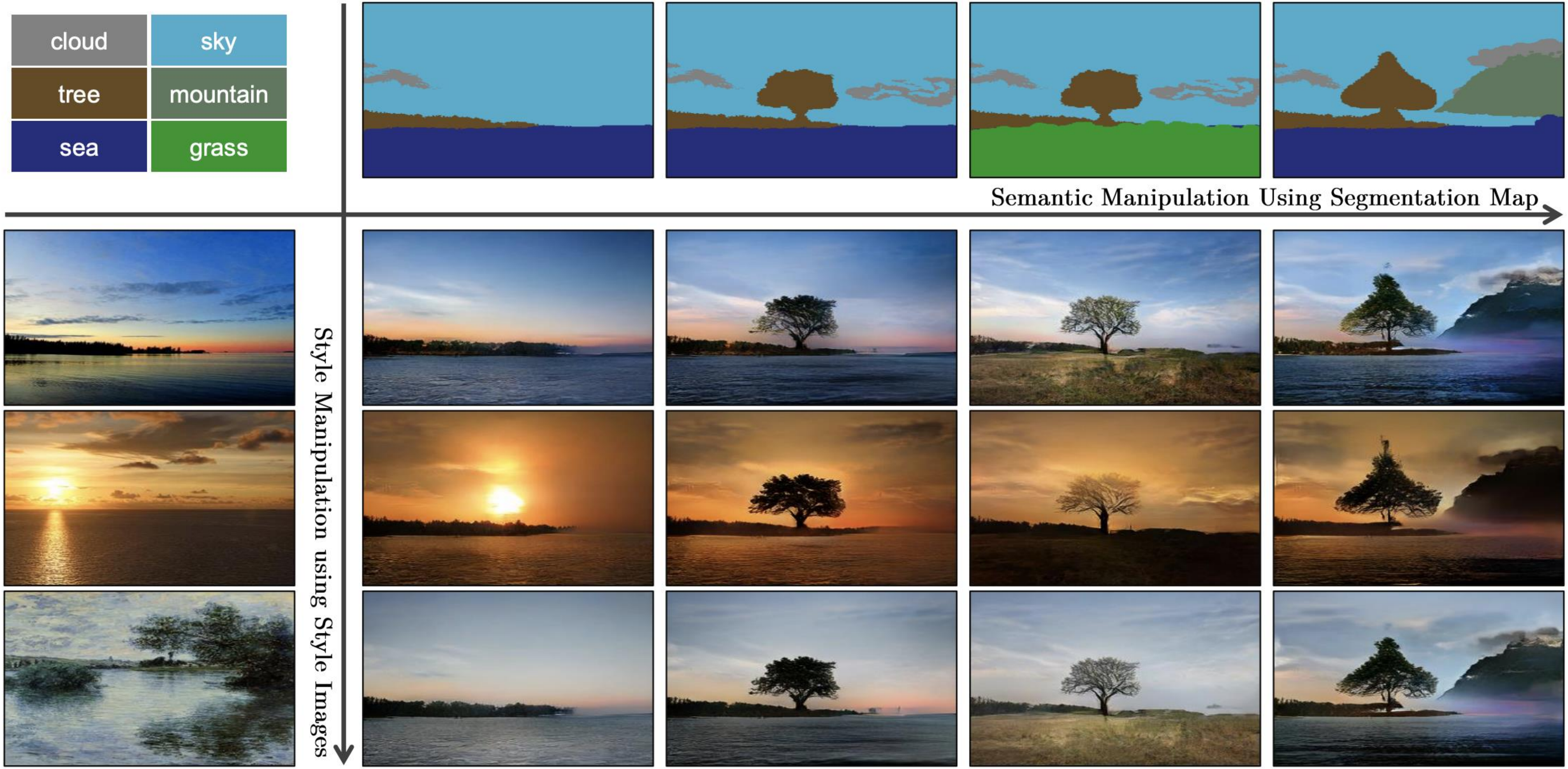
Semantic Image Synthesis



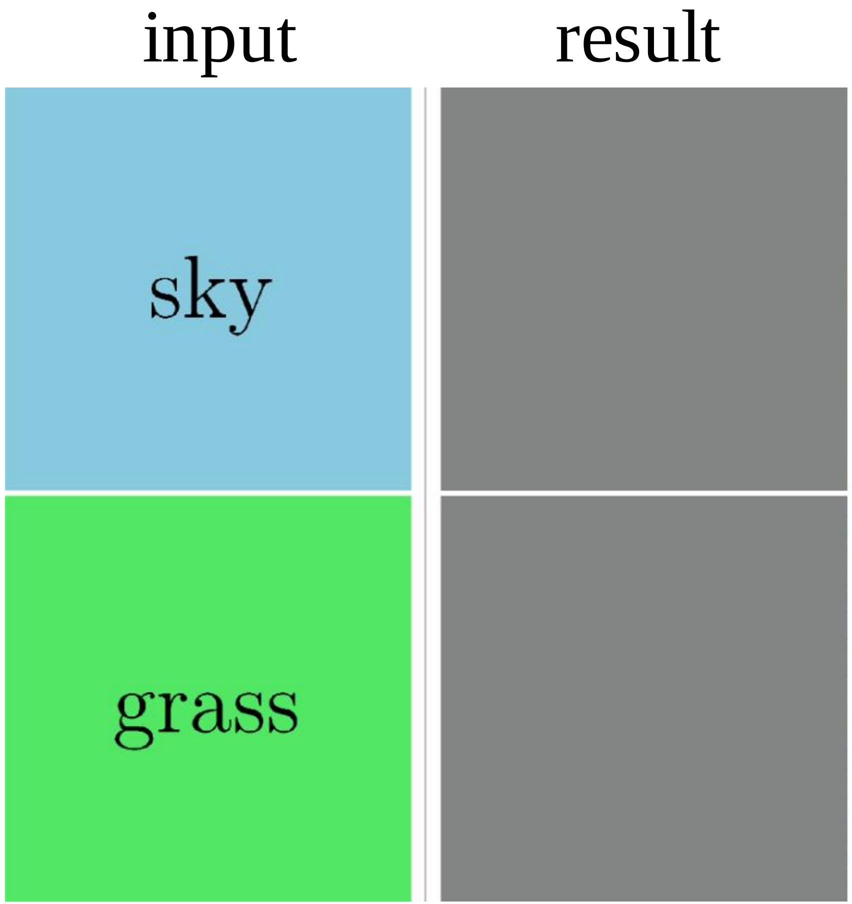
LIVE DEMO

- I need to get an RTX Ready Laptop (https://www.nvidia.com/en-us/geforce/gaming-laptops/20-series/)
- It is running live in GTC
- Will be online for everyone to try out in NVIDIA AI Playground website (https://www.nvidia.com/en-us/research/ai-playground/)


sky
tree
cloud
mountain
grass
sea
river
rock
plant
sand
snow
water
hill
dirt
road
flower
stone
bush
wood
gravel


cloud
sky
tree
mountain
sea
grass
Semantic Manipulation Using Segmentation Map
Style Manipulation using Style Images

PROBLEM WITH PREVIOUS METHODS



PROBLEM WITH PREVIOUS METHODS

Batch Norm (Ioffe et al. 2015)

$x = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ $x = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$

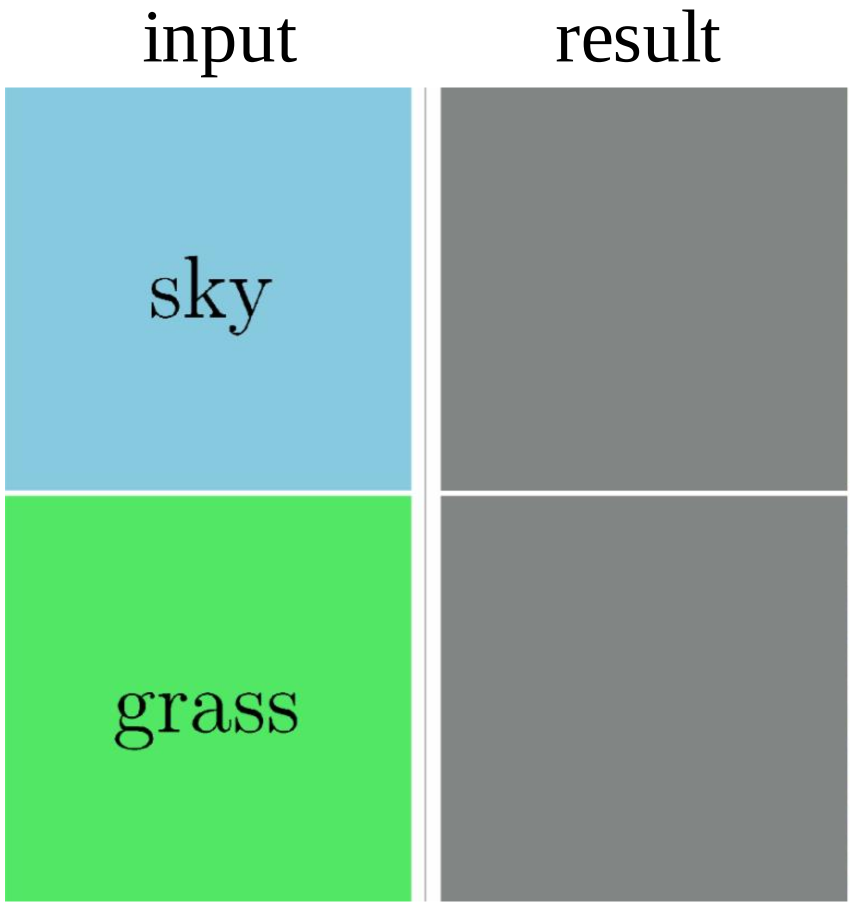
same output!

$$y = \underbrace{\boxed{\frac{x - \mu}{\sigma}}}_{\text{normalization}} \cdot \underbrace{\gamma + \beta}_{\text{affine transform}}$$

affine transform
de-normalization

removes label information

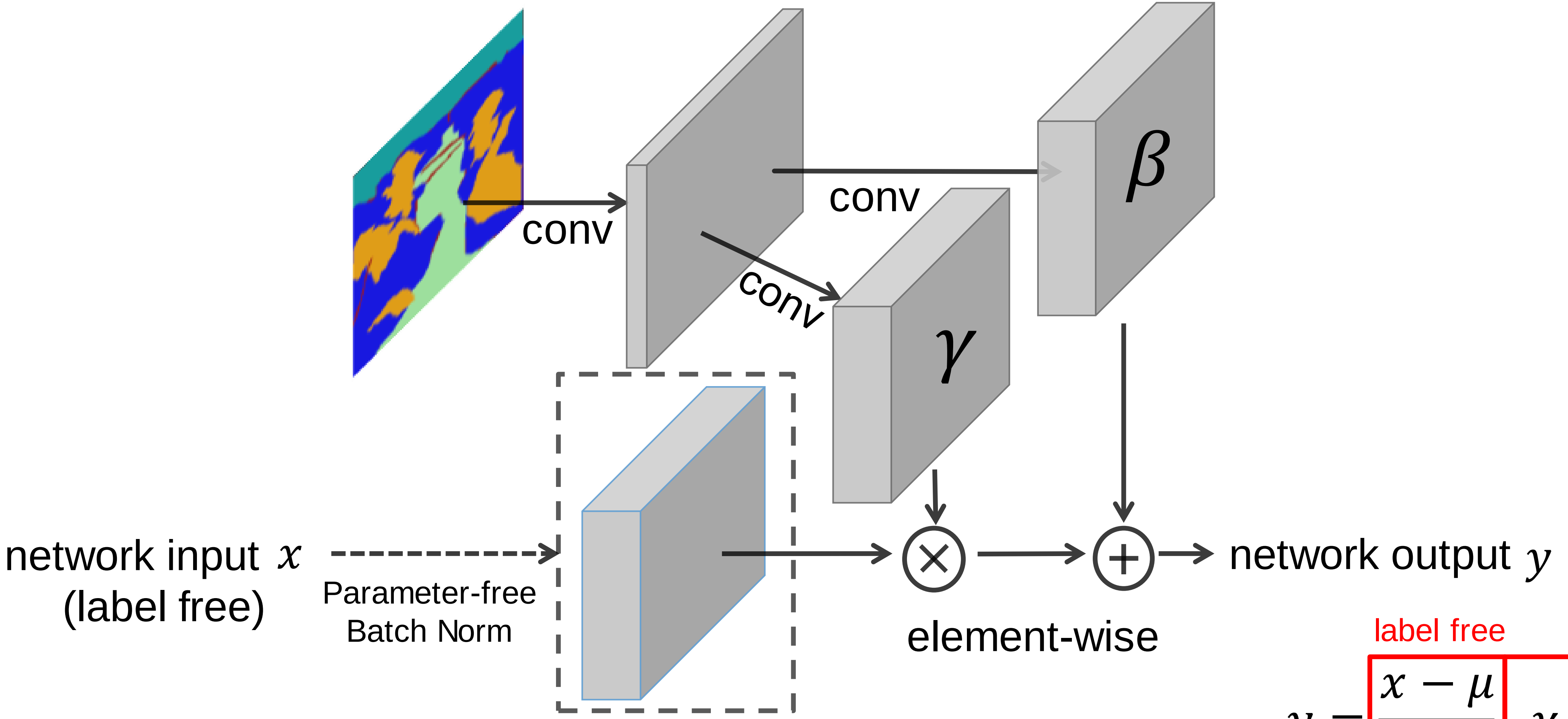
PROBLEM WITH PREVIOUS METHODS



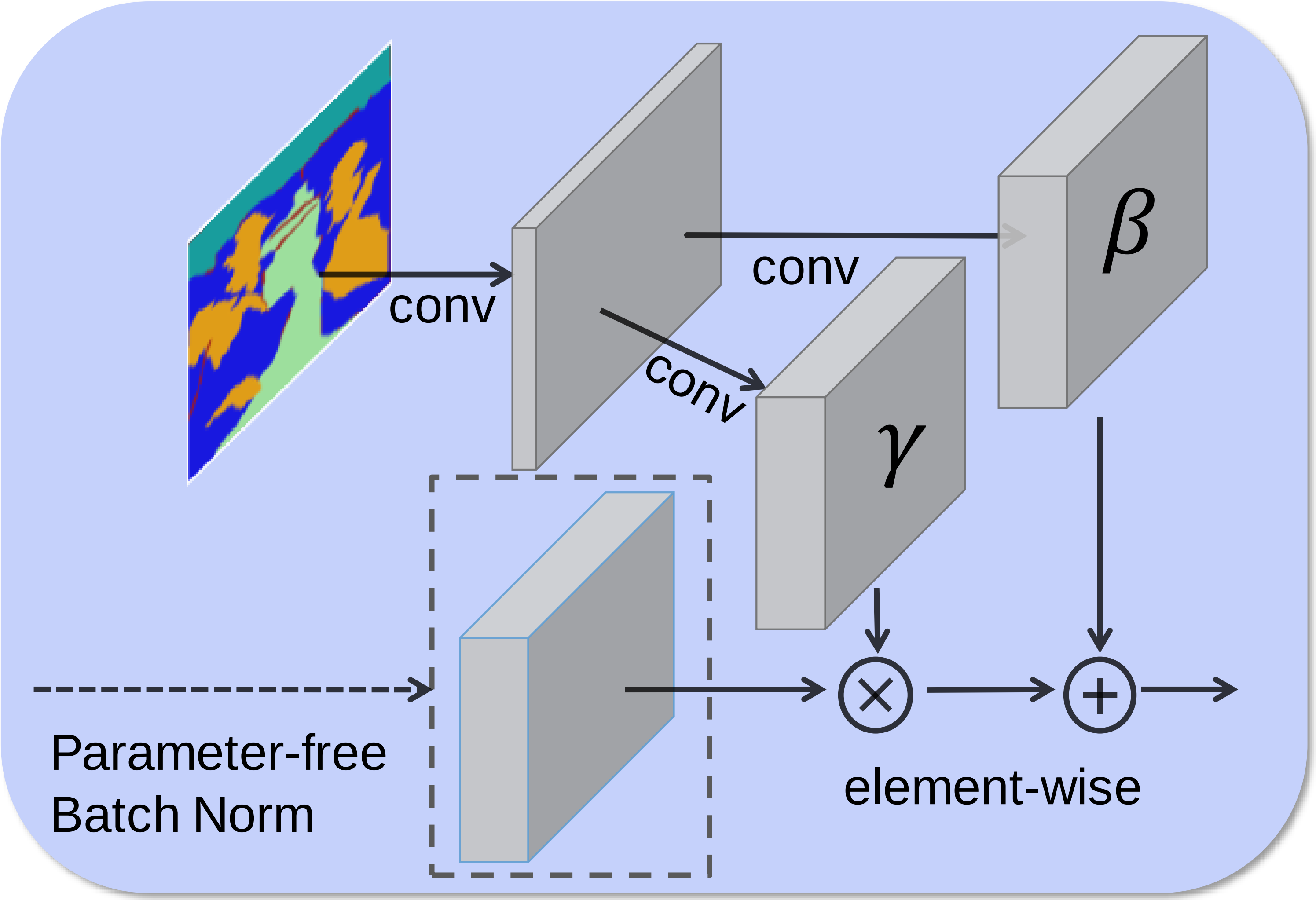
PROBLEM WITH PREVIOUS METHODS

- Do not feed the label map directly to network
- Use the label map to generate normalization layers instead

SPADE(**SP**atially **A**daptive **DE**normalization)



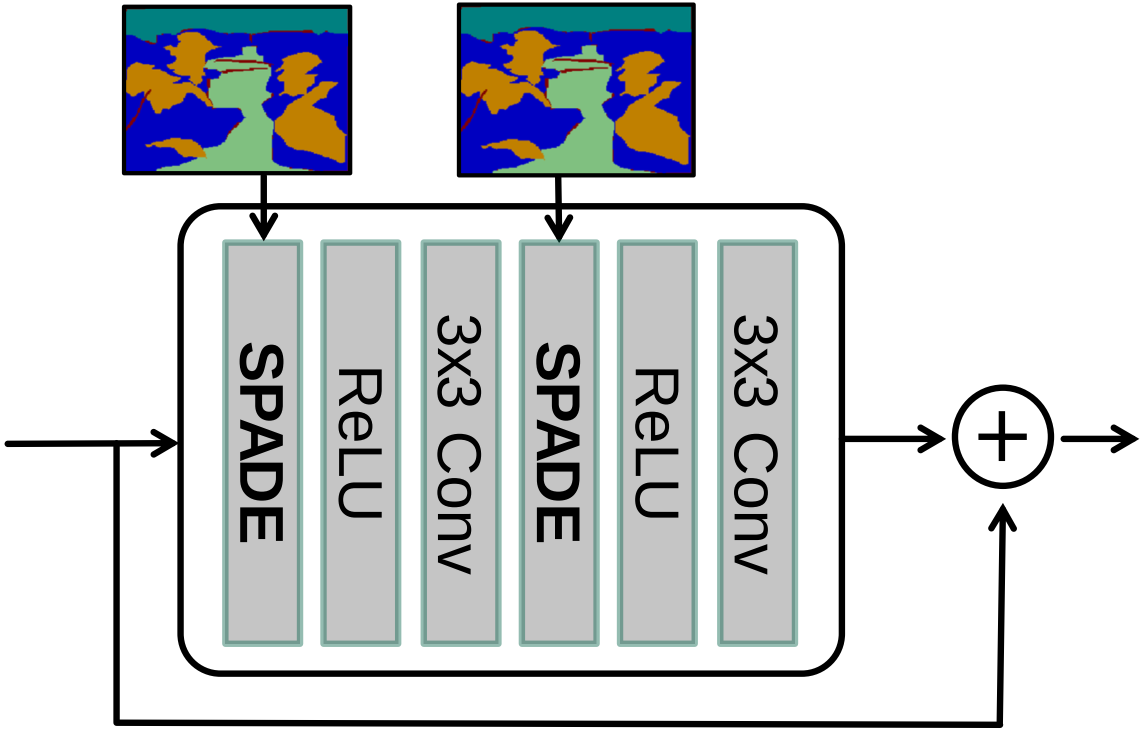
$$y = \boxed{\frac{x-\mu}{\sigma}} \boxed{\cdot \gamma + \beta}$$



SPADE

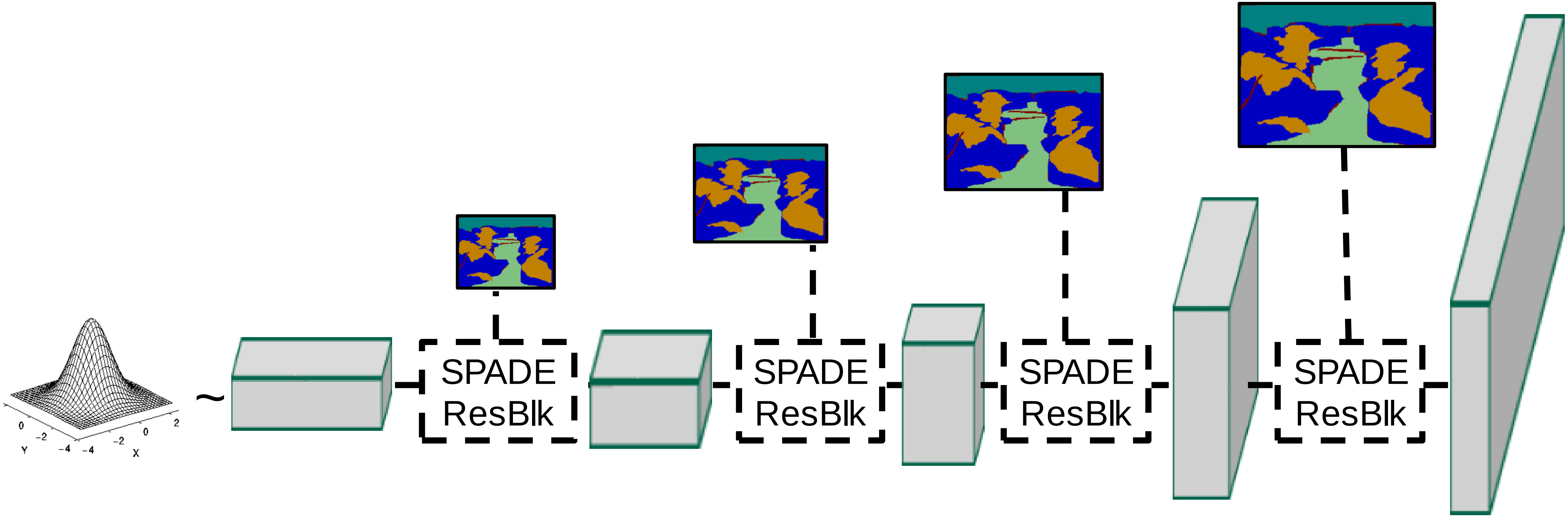
SPatially Adaptive DE-normalization

SPADE RESIDUAL BLOCKS

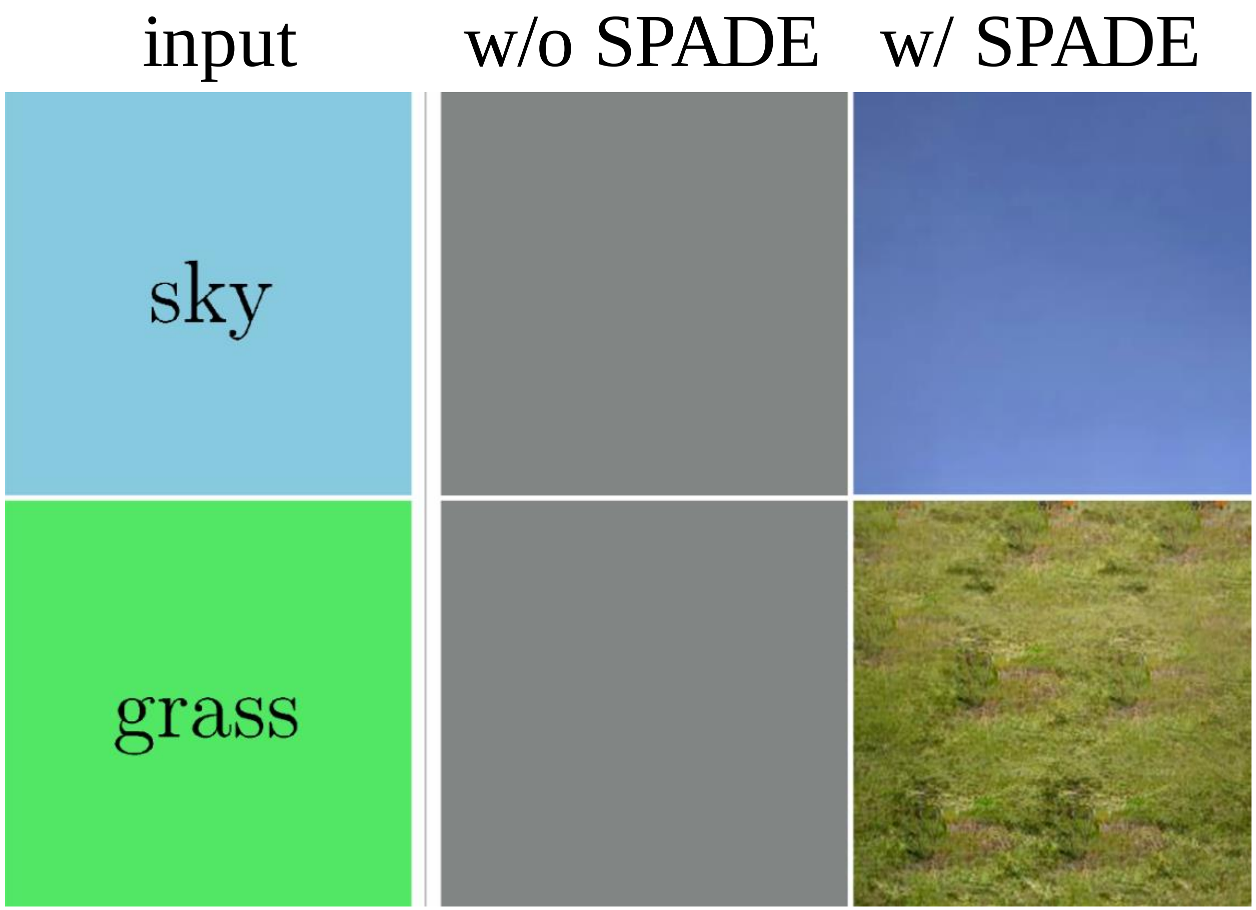


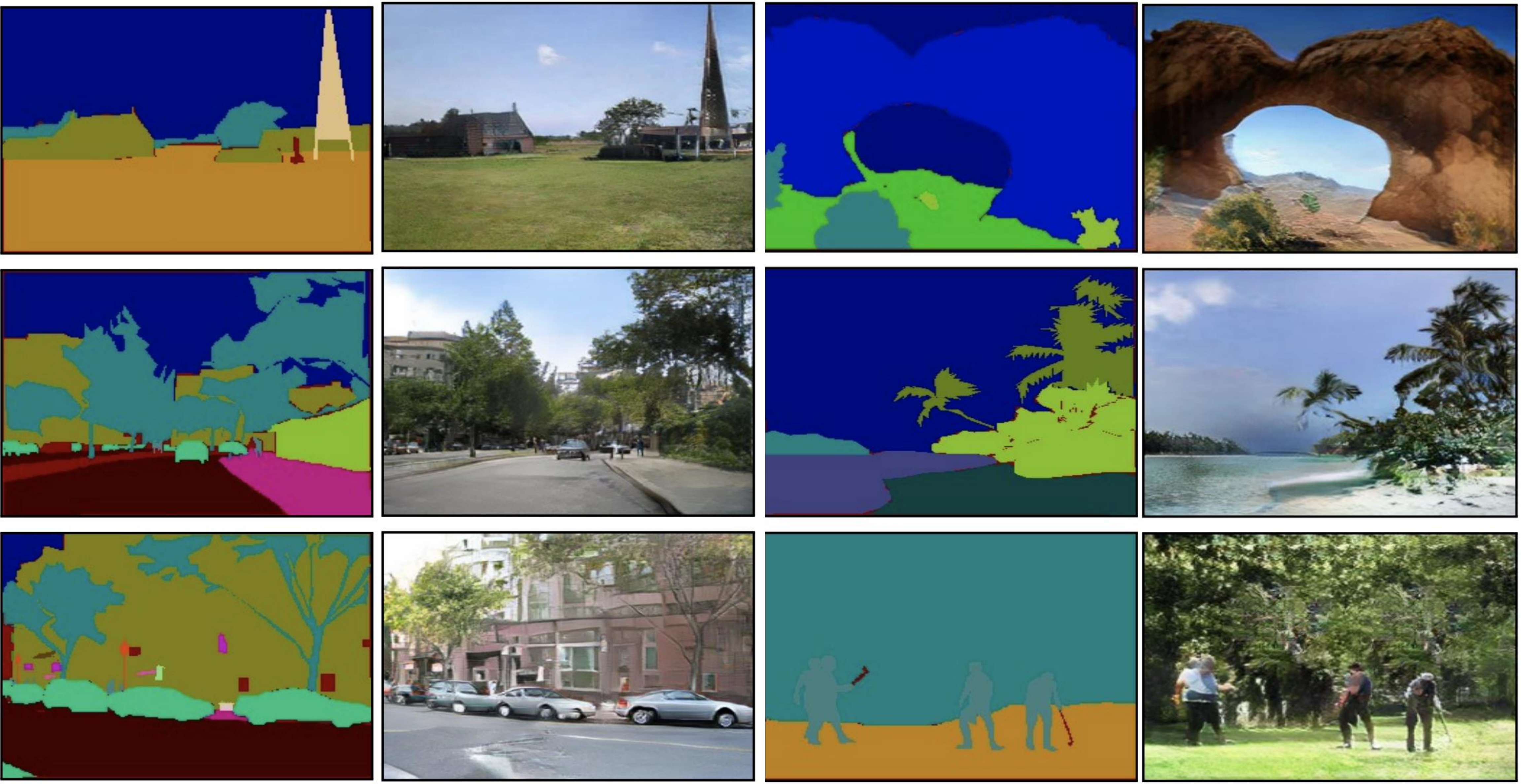
SPADE ResBlk

SPADE GENERATOR



PROBLEM WITH PREVIOUS METHODS





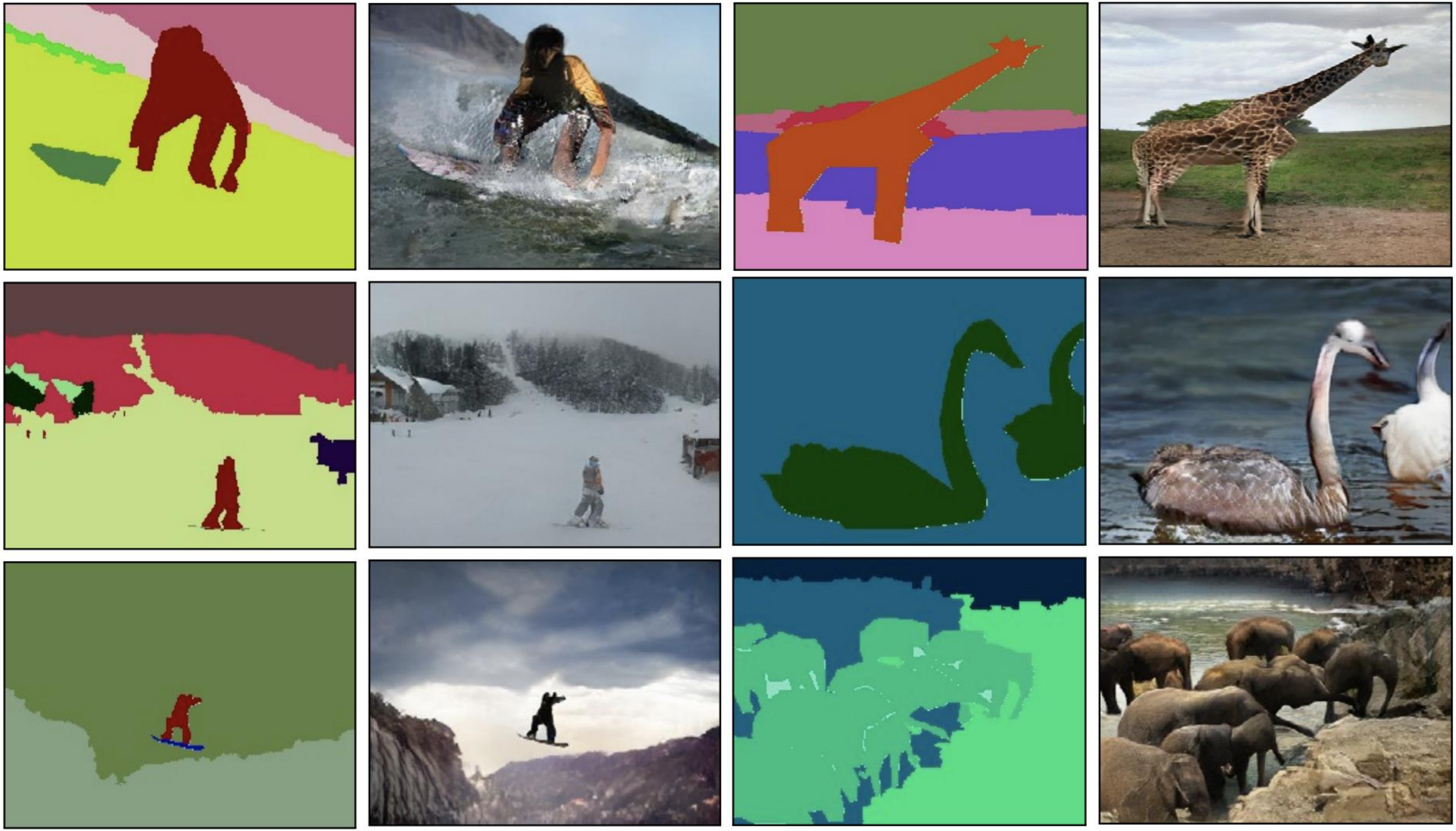
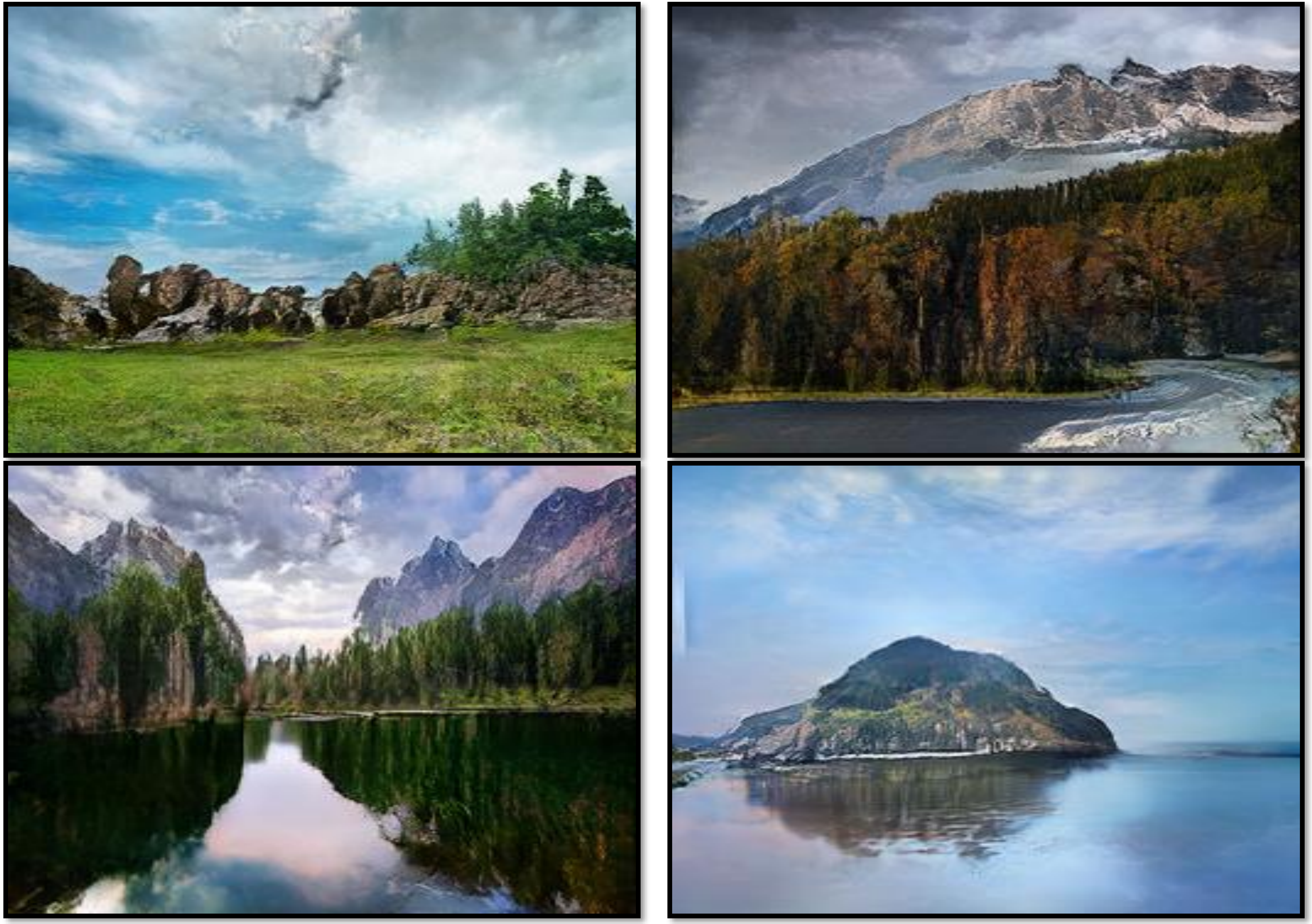
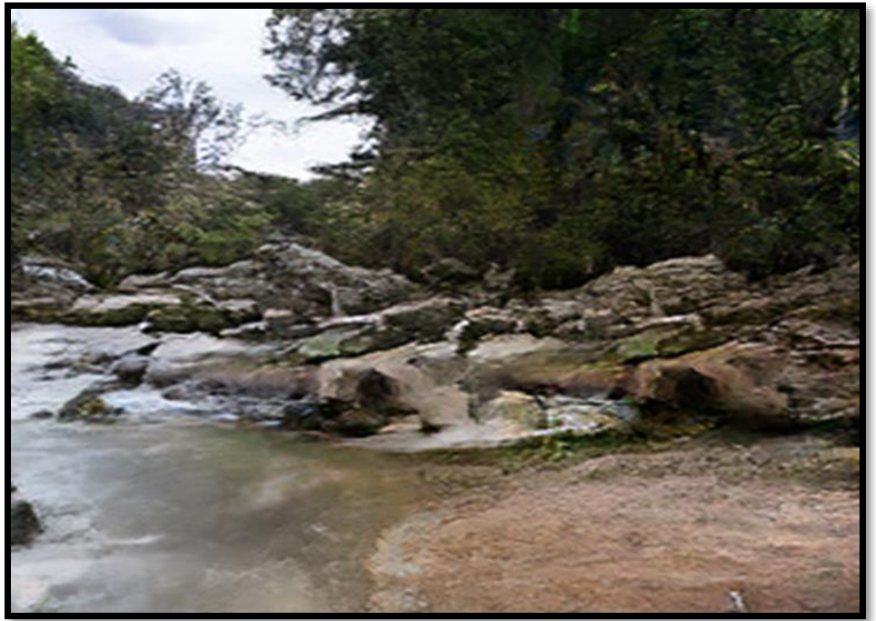


IMAGE RESULTS

IMAGE RESULTS



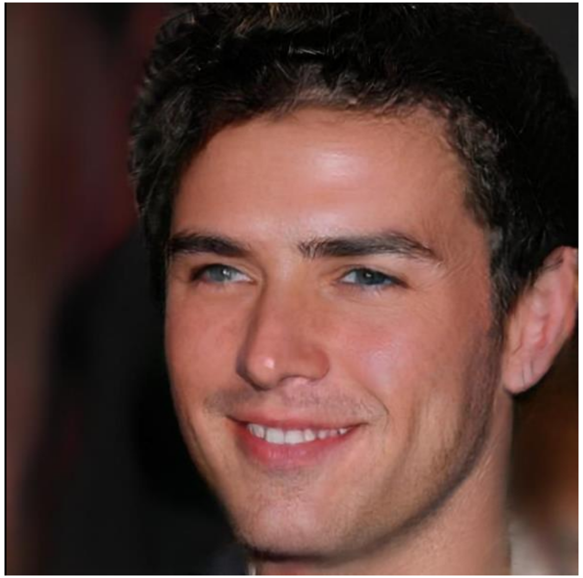




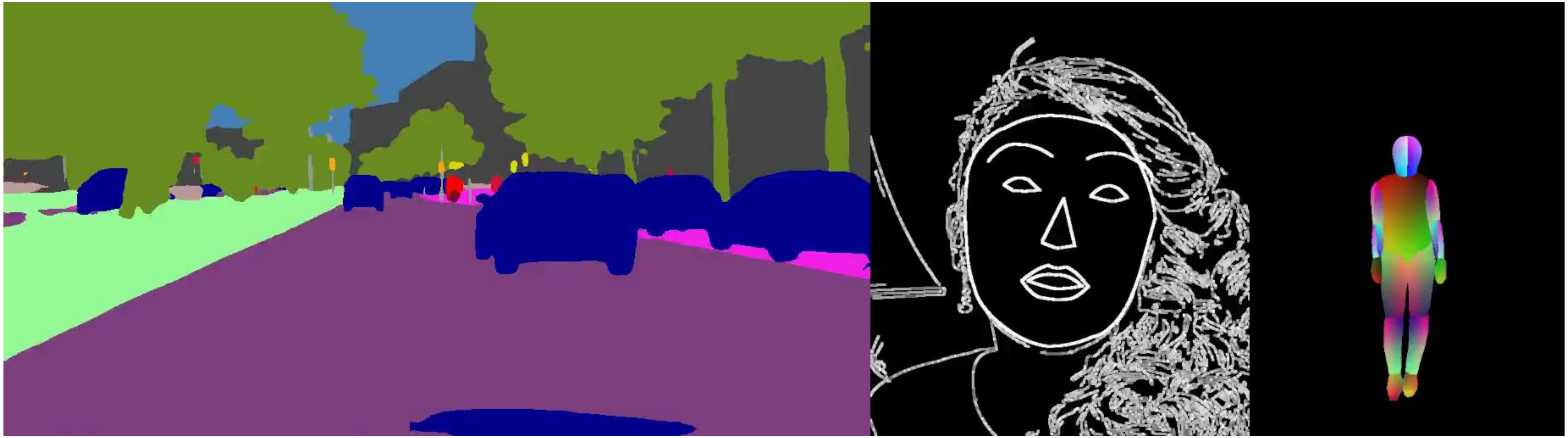


VIDEO-TO-VIDEO SYNTHESIS

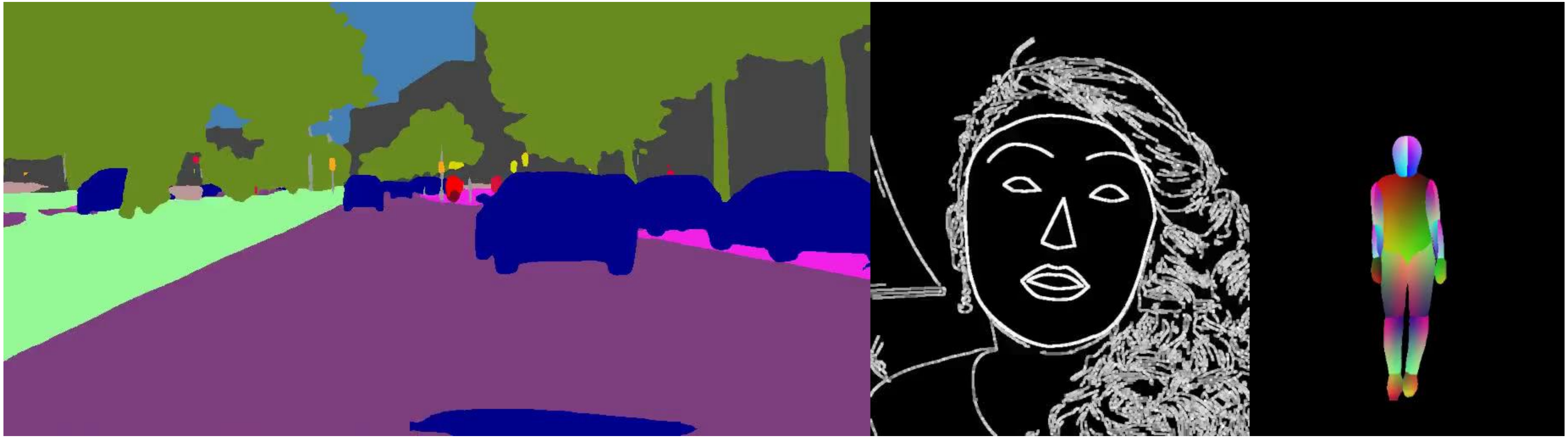
IMAGE-TO-IMAGE SYNTHESIS



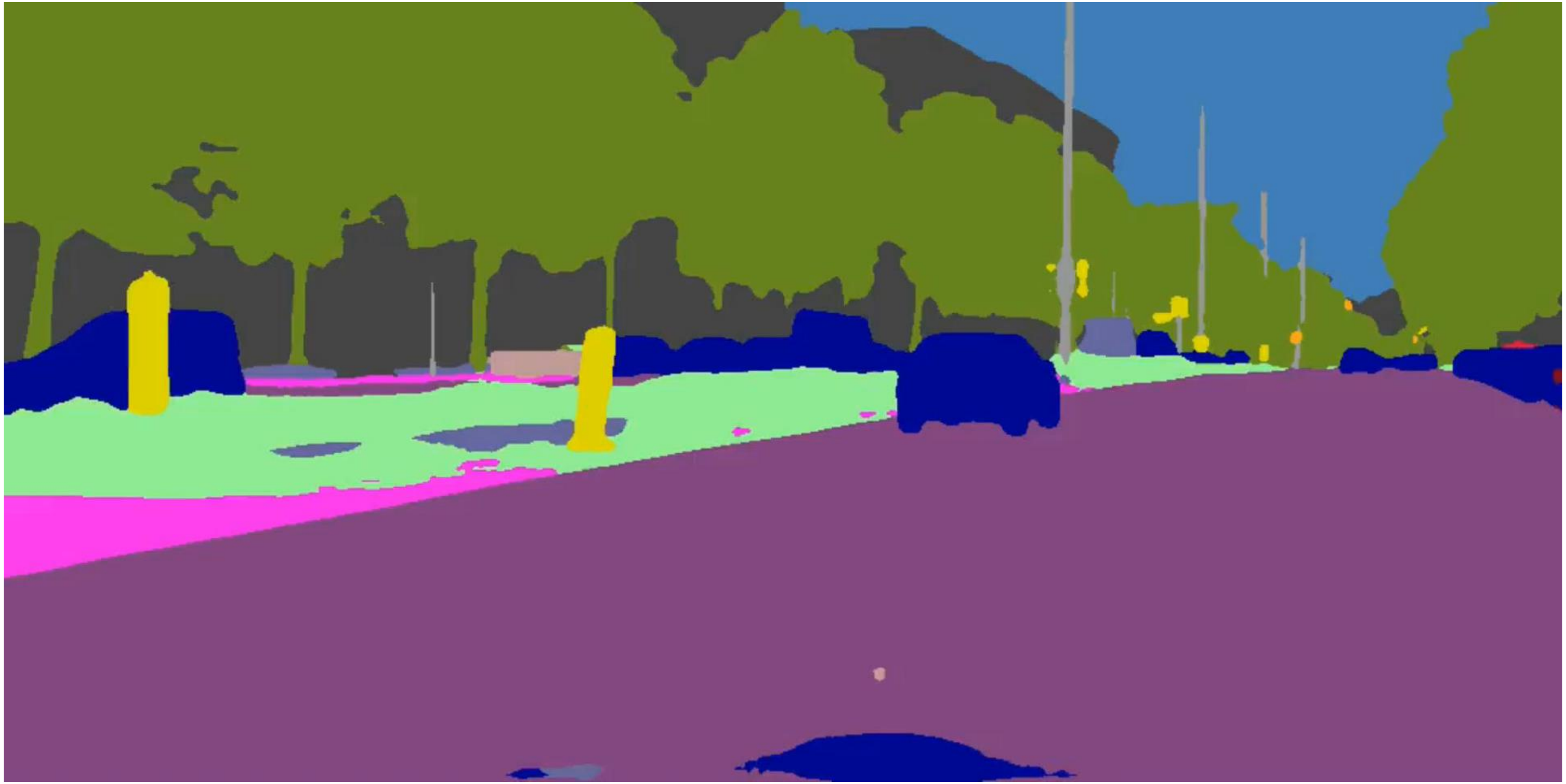
VIDEO-TO-VIDEO SYNTHESIS



VIDEO-TO-VIDEO SYNTHESIS



VIDEO-TO-VIDEO SYNTHESIS



VIDEO-TO-VIDEO SYNTHESIS

MOTIVATION

- AI-based rendering

Traditional graphics

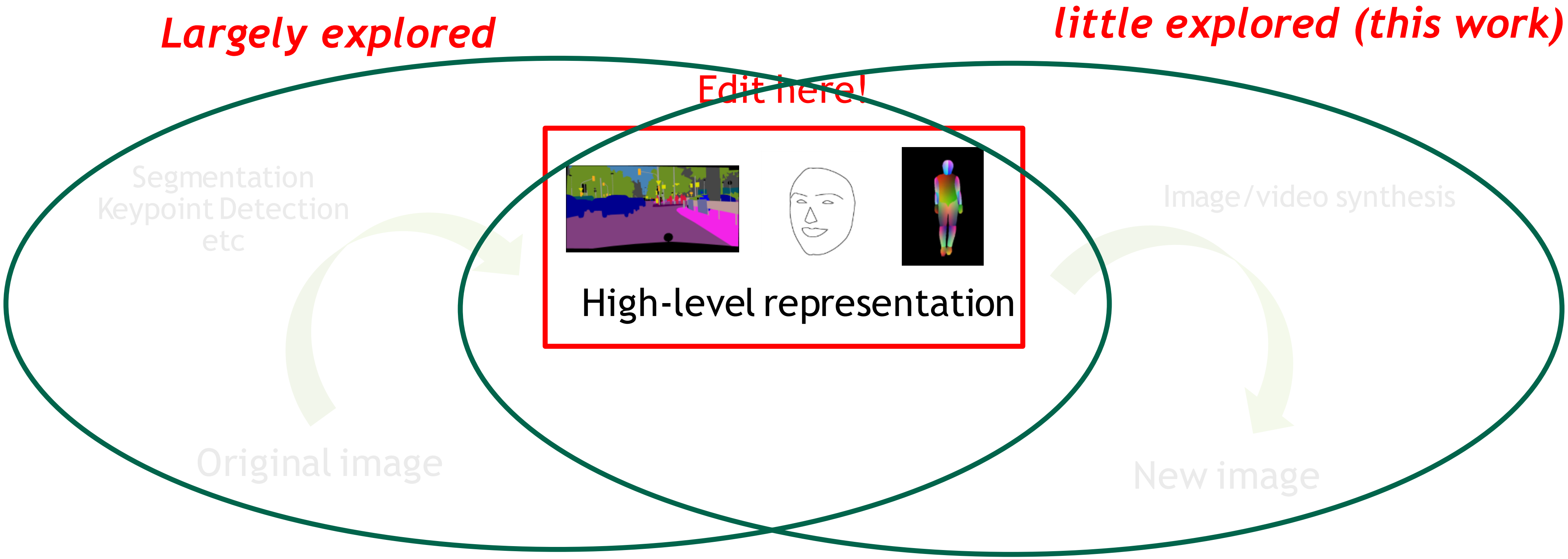
Geometry, texture, lighting

Machine learning graphics

Data

MOTIVATION

- AI-based rendering
- High-level semantic manipulation



PREVIOUS WORK

Image translation

pix2pixHD [2018], CRN [2017], pix2pix [2017]

Unconditional synthesis

MoCoGAN [2018], TGAN [2017], VGAN [2016]

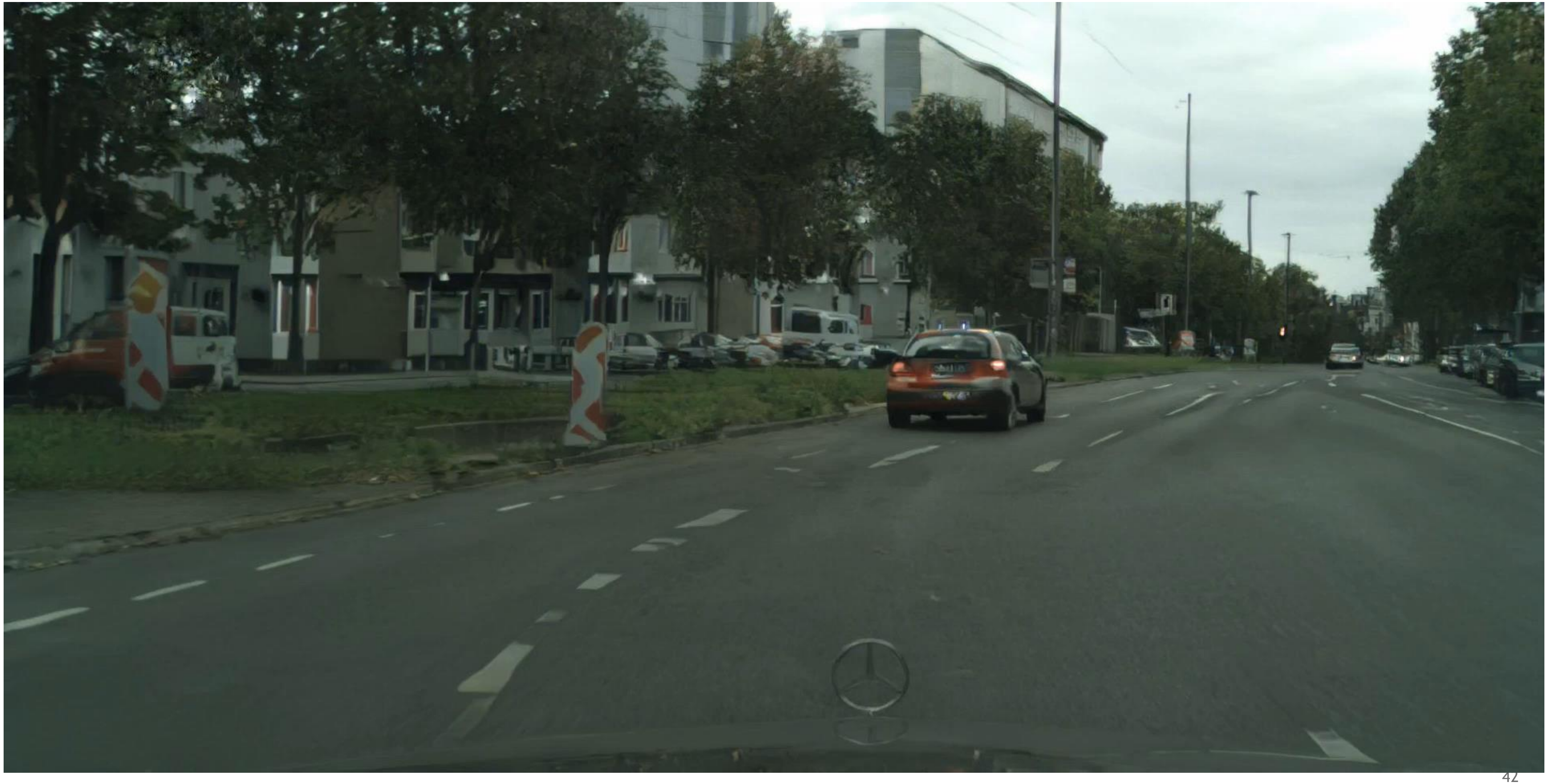
Video style transfer

COVST [2017], ArtST [2016]

Video prediction

MCNet [2017], PredNet [2017]

PREVIOUS WORK: FRAME-BY-FRAME RESULT

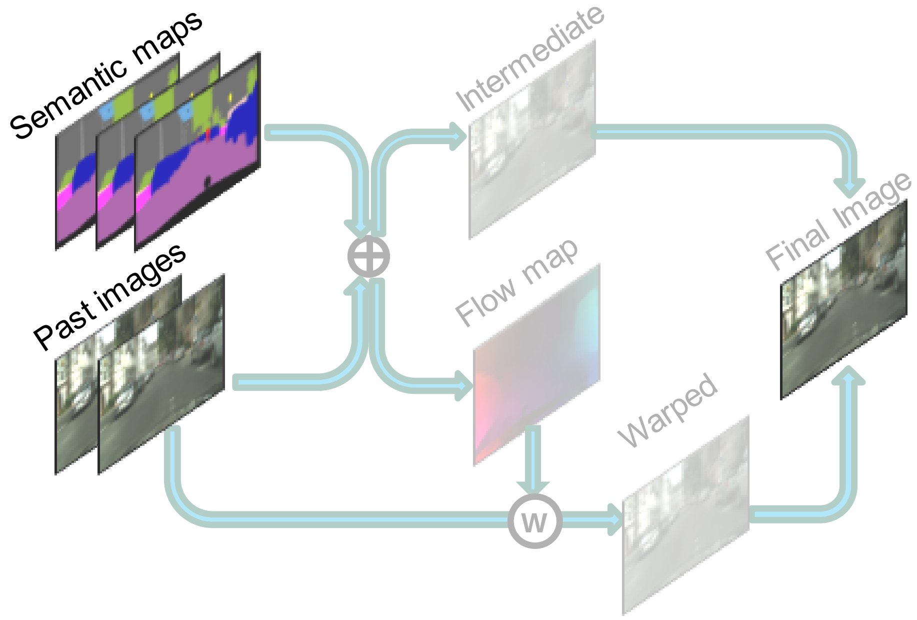


OUR METHOD

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

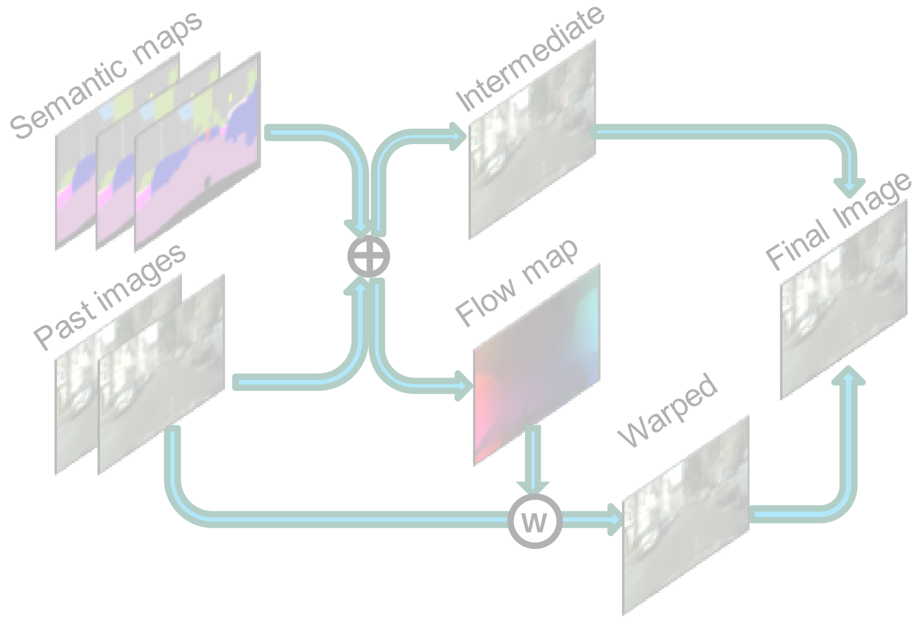
OUR METHOD

Sequential Generator



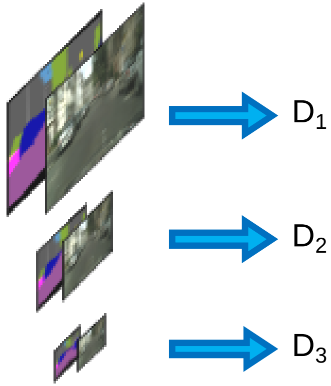
OUR METHOD

Sequential Generator

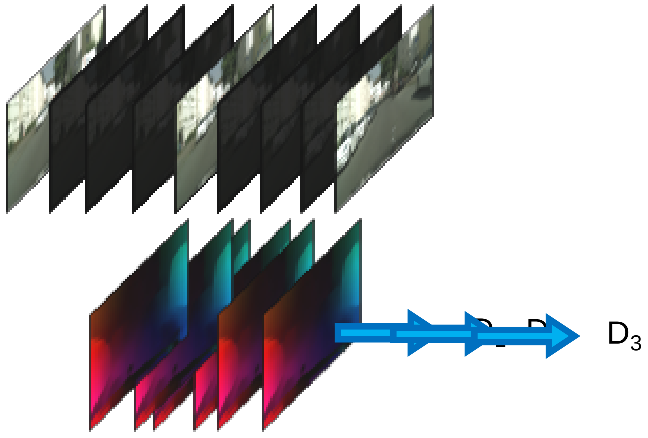


Multi-scale Discriminators

Image Discriminator



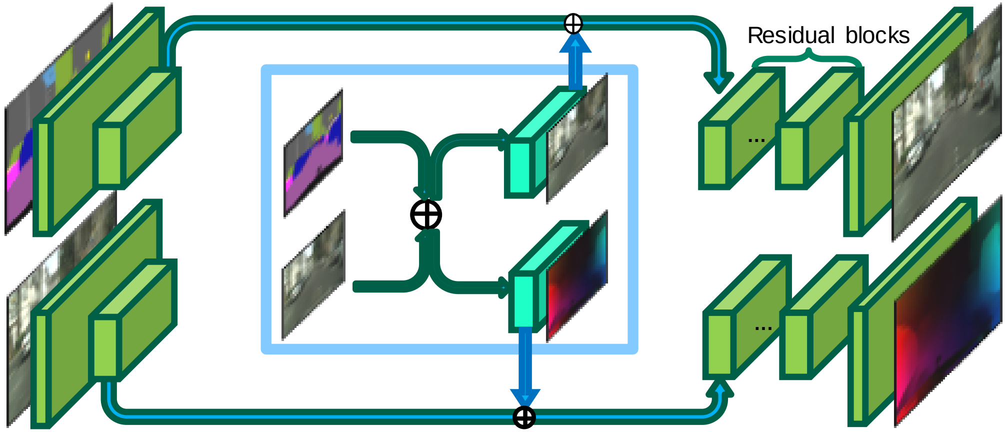
Video Discriminator



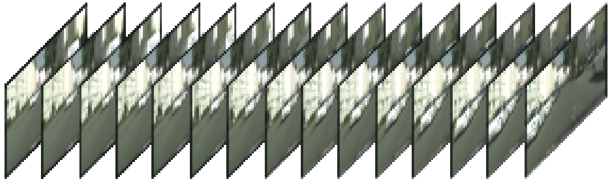
OUR METHOD

Spatio-temporally Progressive Training

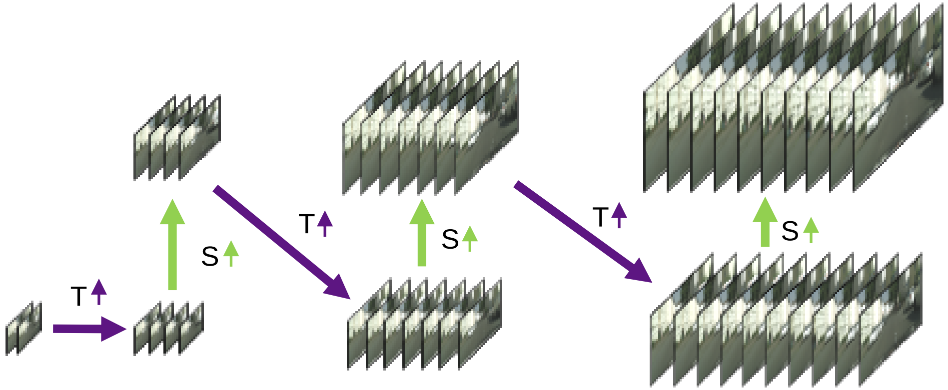
Spatially progressive



Temporally progressive



Alternating training



RESULTS

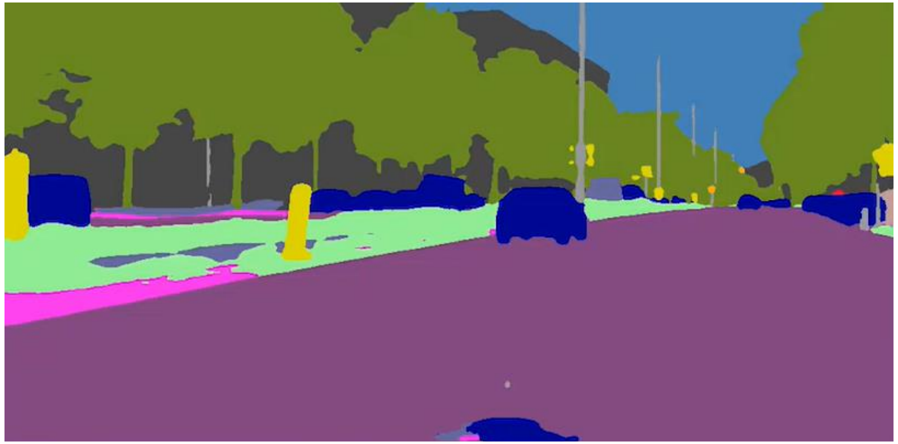
RESULTS

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

RESULTS

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

STREET VIEW: CITYSCAPES



Semantic map

pix2pixHD

COVST (video style transfer)

Ours

STREET VIEW: BOSTON

STREET VIEW: NYC

RESULTS

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

FACE SWAPPING (FACE → EDGE → FACE)

input edges output

FACE SWAPPING (SLIMMER FACE)

input

(slimmed) edges

(slimmed) output

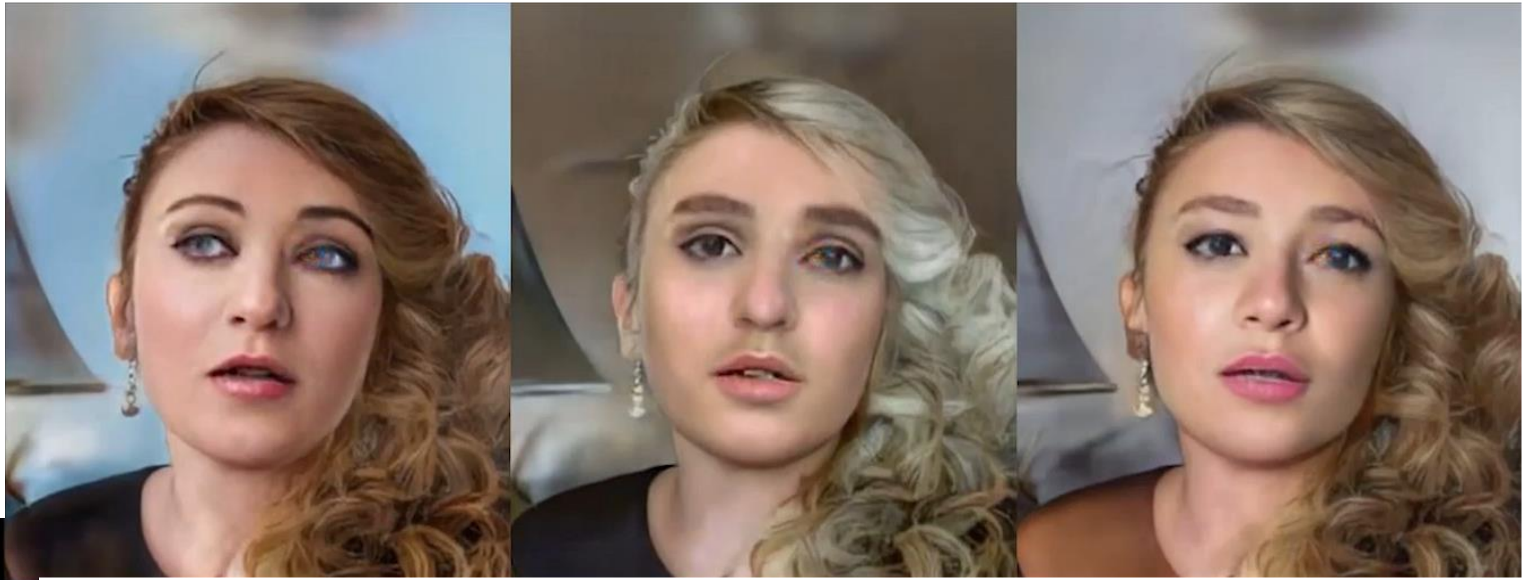
FACE SWAPPING (SLIMMER FACE)

input

(slimmed) edges

(slimmed) output

MULTI-MODAL EDGE → FACE



Style 1 Style 2 Style 3

RESULTS

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

NVIDIA

MOTION TRANSFER (BODY → POSE → BODY)

input poses output

MOTION TRANSFER

EXTENSION: FRAME PREDICTION

- Goal: predict future frames given past frames
- Our method: decompose prediction into two steps
    - 1. predict the semantic map for next frame
    - 2. synthesize the frame based on the semantic map

EXTENSION: FRAME PREDICTION

Ground truth

PredNet

MCNet

Ours

INTERACTIVE GRAPHICS

PATH TO INTERACTIVE GRAPHICS

- Real-time inference
- Combining with existing graphics pipeline
- Domain gap between real input and synthetic input

PATH TO INTERACTIVE GRAPHICS

- Real-time inference
- Combining with existing graphics pipeline
- Domain gap between real input and synthetic input

PATH TO INTERACTIVE GRAPHICS

- Real-time inference
    - FP16 + TensorRT → ~5 times speed up
    - 36ms (27.8 fps) for 1080p inference
    - Overall: 15~25 fps

PATH TO INTERACTIVE GRAPHICS

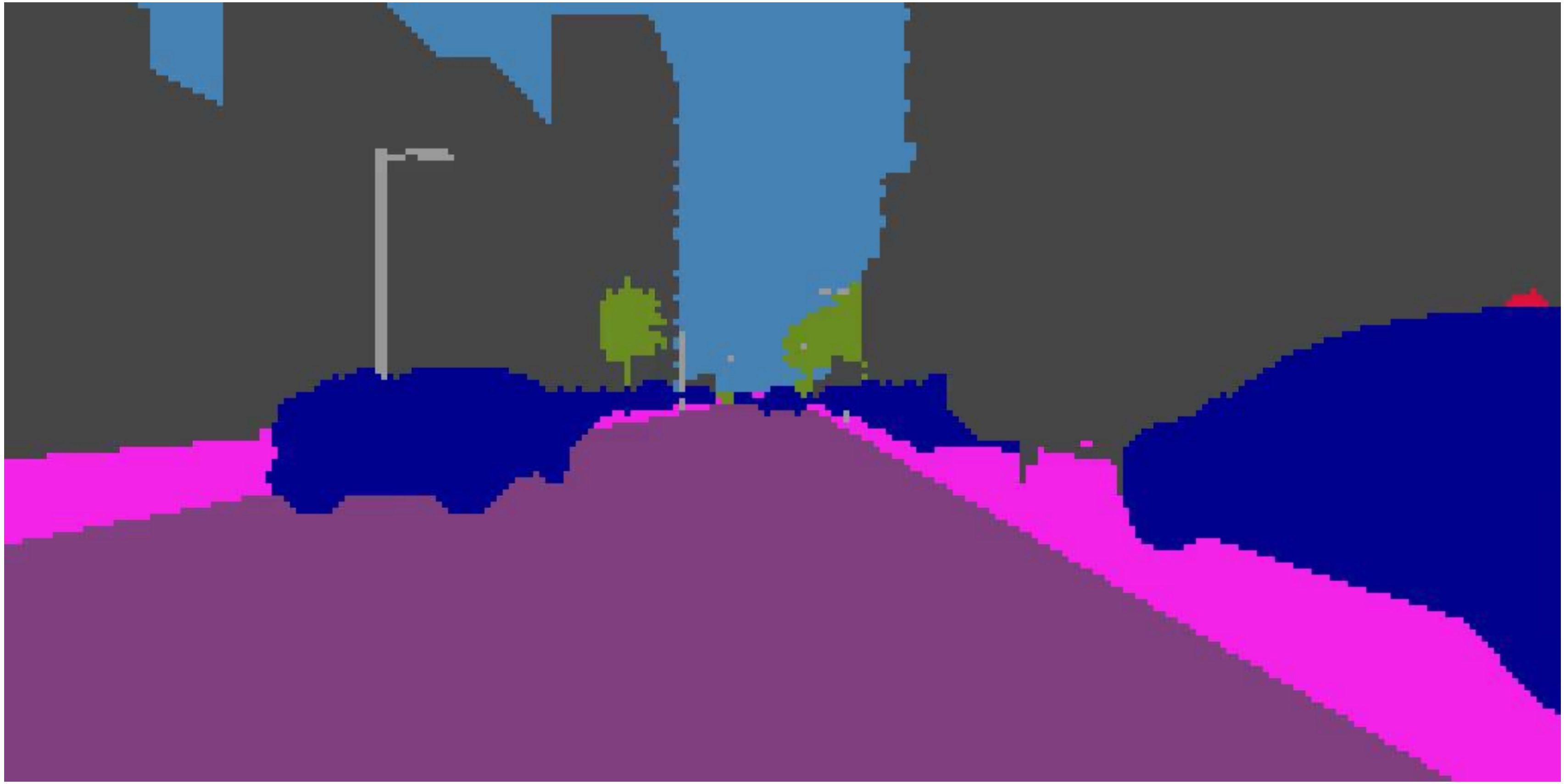
- Real-time inference
- Combining with existing graphics pipeline
  - CARLA: open-source simulator for autonomous driving research
  - Make game engine render semantic maps
  - Pass the maps to the network and display the inference result

PATH TO INTERACTIVE GRAPHICS

- Real-time inference
- Combining with existing graphics pipeline
- Domain gap between ***real*** input and ***synthetic*** input
    - Network trained on real data but tested on synthetic data
    - Things that differ: Object shapes/edges, density of objects, camera viewpoints, etc
    - On-going work

ORIGINAL CARLA IMAGE

RENDERED SEMANTIC MAPS



RECORDED DEMO RESULTS

RECORDED DEMO RESULTS

CONCLUSION

CONCLUSION

- What can we achieve?
- What can it be used for?

CONCLUSION

- What can we achieve?
    - Synthesize high-res realistic images

CONCLUSION

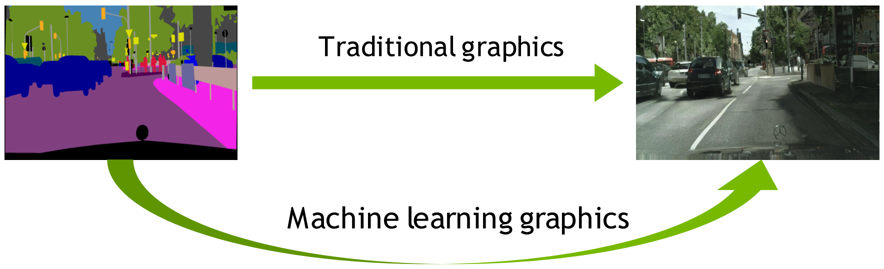
- What can we achieve?
  - Synthesize high-res realistic images
  - Produce temporally-smooth videos

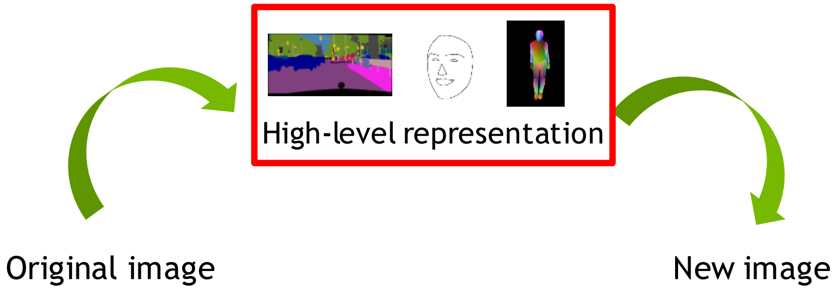
CONCLUSION

- What can we achieve?
  - Synthesize high-res realistic images
  - Produce temporally-smooth videos
  - Reinvent interactive graphics

CONCLUSION

- What can we achieve?
- What can it be used for?
    - AI-based rendering
    - High-level semantic manipulation





THANK YOU

https://github.com/NVIDIA/vid2vid