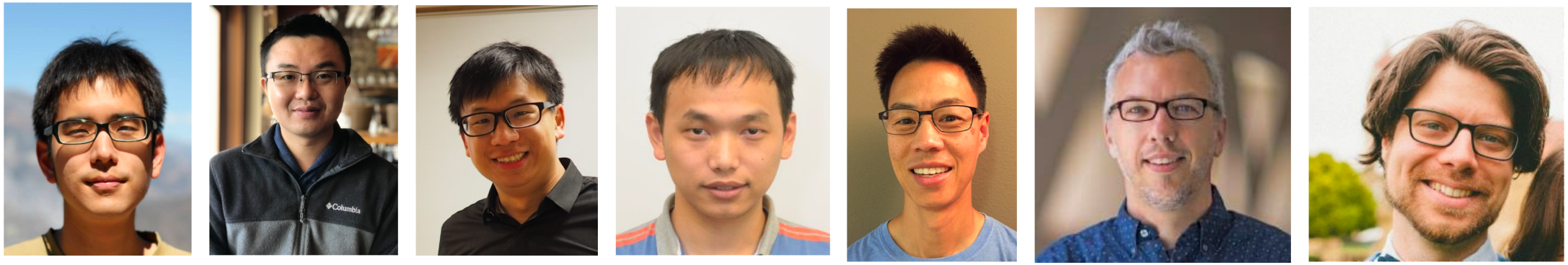
Video-to-Video Synthesis

Ming-Yu Liu
NVIDIA



Outline

- Introduction
- Method
- Results
- Next Frame Prediction
- Conclusion

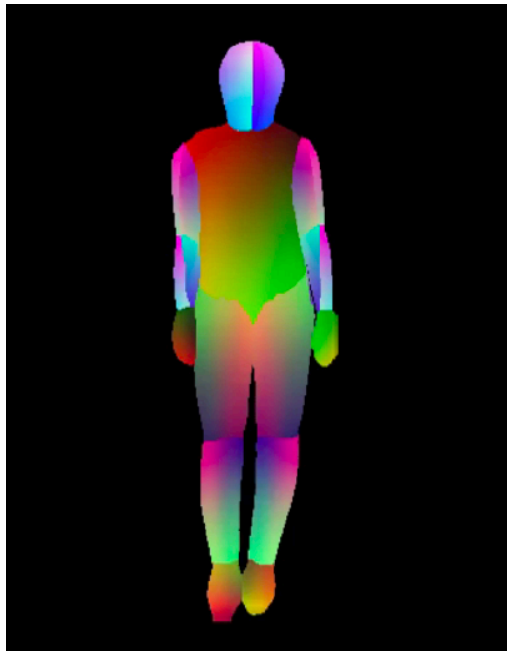
Outline

- **Introduction**
- Method
- Results
- Next Frame Prediction
- Conclusion

The Video-to-Video Synthesis Problem

Sequence of
Semantic Representations

Photorealistic Video



Introduction - Why

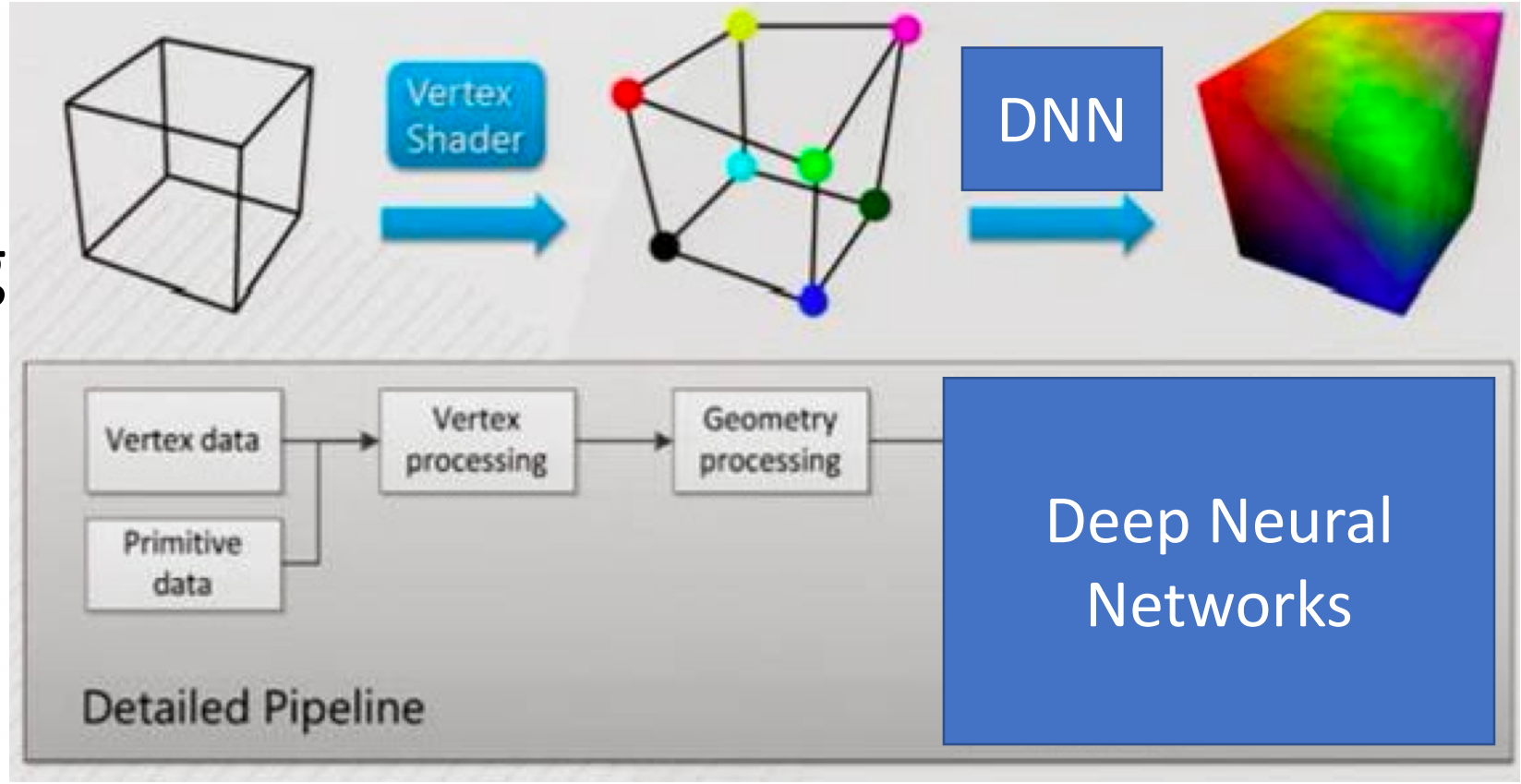
- Deep Imagination
  - The Mind's Eye

"You Can Do Anything in Your Minds Eye"

Image credit: https://medium.com/thrive-global/make-your-imagination-work-for-you-49f8be368965

Introduction - Why

- Deep Imagination
  - The Mind's Eye
- Alternative Way for CG Rendering



http://3dthebasics.blogspot.com/2013/01/ha7-task-2-displaying-3d-polygon.html

Introduction - Why

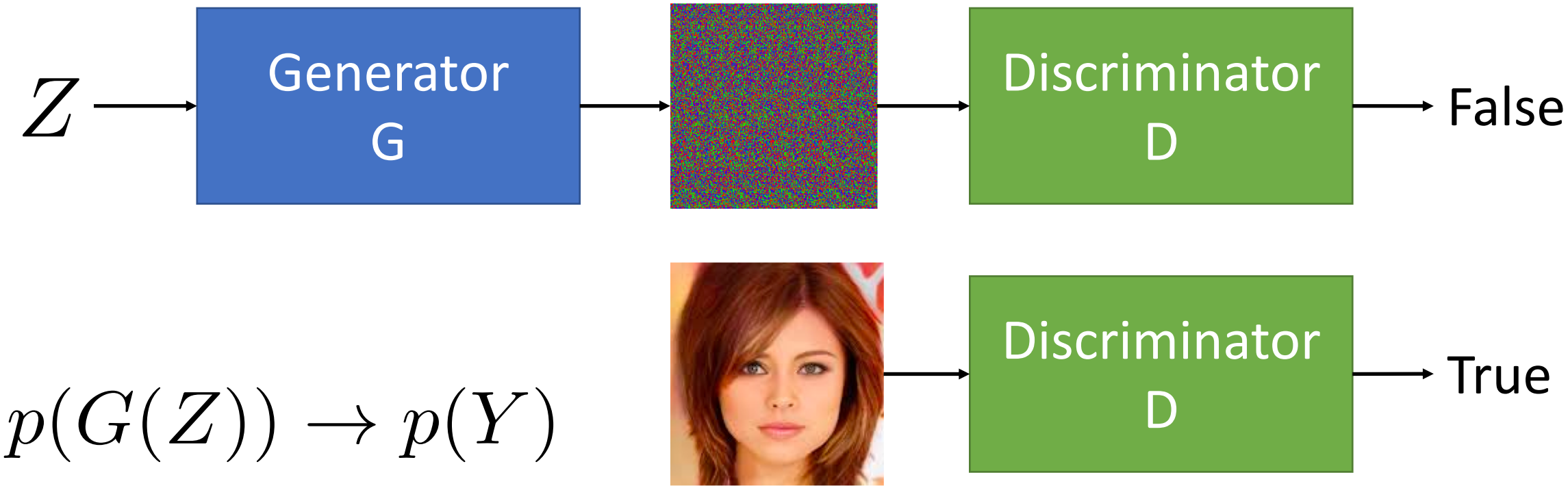
- Deep Imagination
  - The Mind's Eye
- Alternative Way for CG Rendering
- Model-based Reinforcement Learning

World Model, Ha and Schmidhuber, arxiv 2018

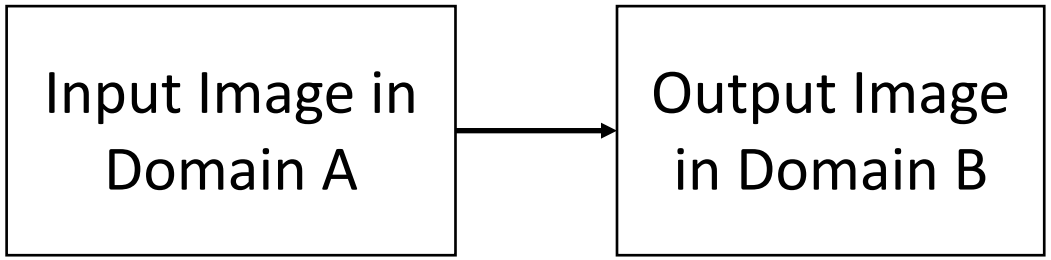
Introduction – Other Applications

- Motion Retargeting

Introduction - GANs



Introduction – Image-to-Image Translation

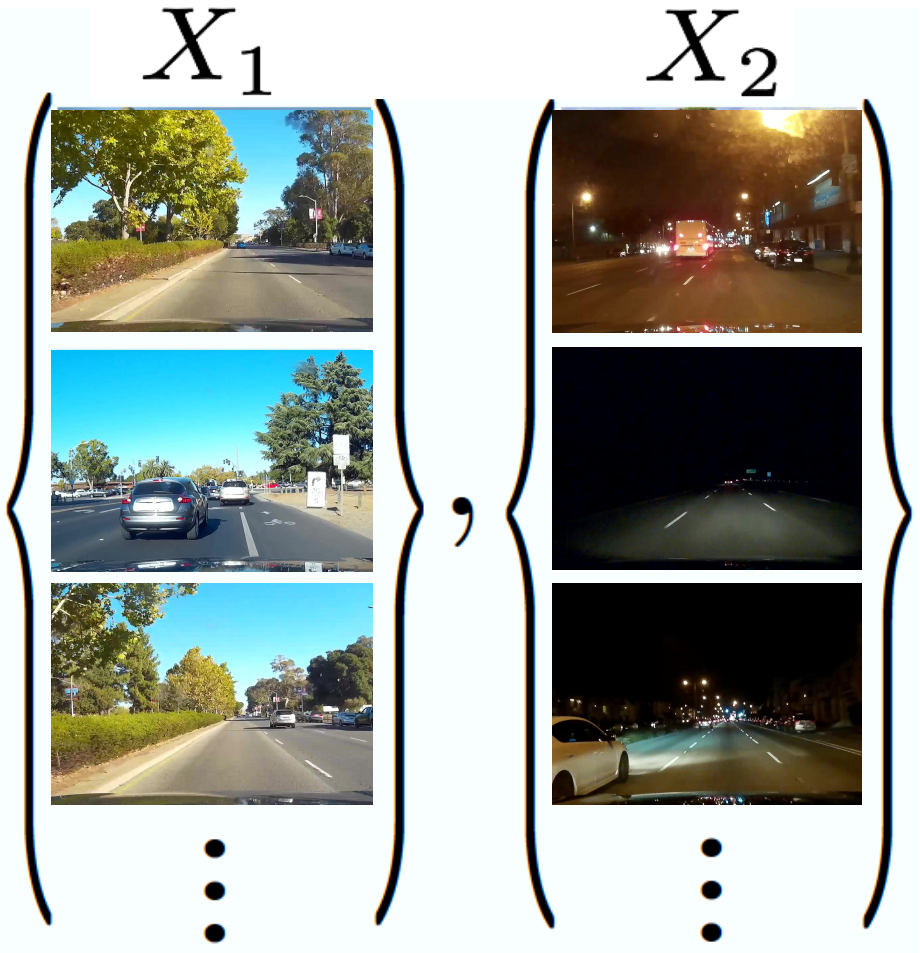


	Supervised	Unsupervised
Unimodal	pix2pix, CRN, SRGAN, … **vid2vid**	SimGAN, DiscoGAN, CycleGAN, UNIT, DTN, DualGAN, StarGAN, XGAN, RecycleGAN, CoupledGAN, …
Multimodal	pix2pixHD, BicycleGAN, SIMS, …	MUNIT, Augmented CycleGAN, DRIT, EG-UNIT, …

Introduction - Supervised vs Unsupervised

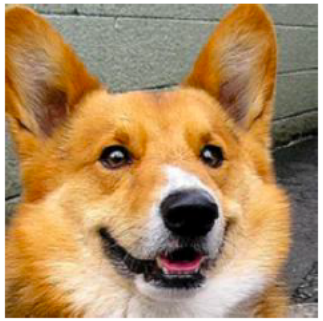
Supervised

Unsupervised

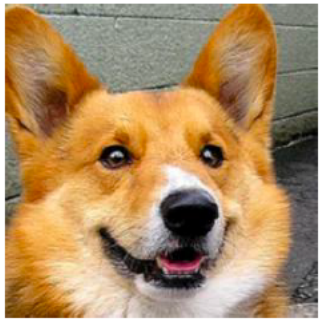
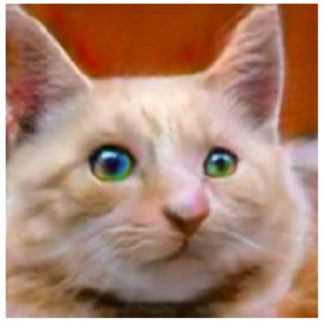


Introduction - Unimodal vs Multimodal

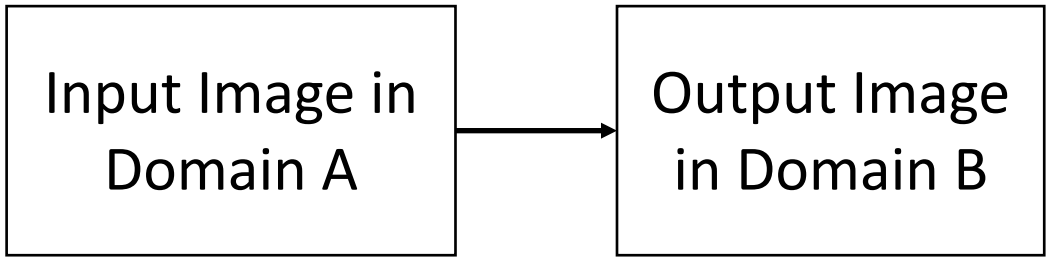
Unimodal $p(Y|X) = \delta(F(X))$

$F($  $) =$ 

Multimodal $p(Y|X) = F(X, S)$

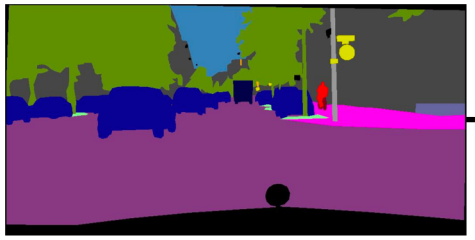
$F($  $) =$   

Introduction – Image-to-Image Translation



	Supervised	Unsupervised
Unimodal	pix2pix, CRN, SRGAN, … **vid2vid**	SimGAN, DiscoGAN, CycleGAN, UNIT, DTN, DualGAN, StarGAN, XGAN, RecycleGAN, CoupledGAN, …
Multimodal	pix2pixHD, BicycleGAN, SIMS, …	MUNIT, Augmented CycleGAN, DRIT, EG-UNIT, …

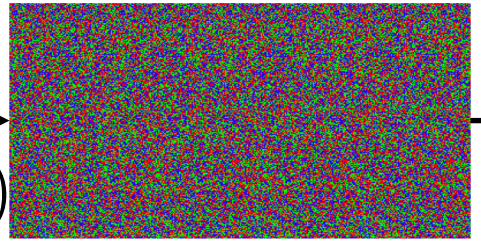
Introduction – Image Conditional GANs

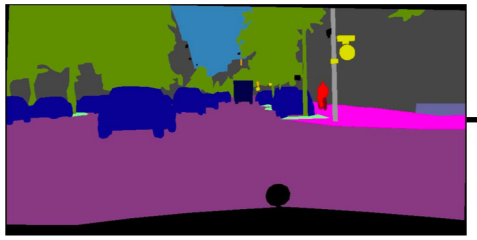


X

Generator

$G(X)$





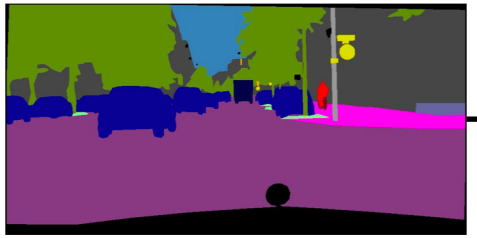
X

Discriminator

False

$p(G(X), X) \rightarrow p(Y, X)$

Y

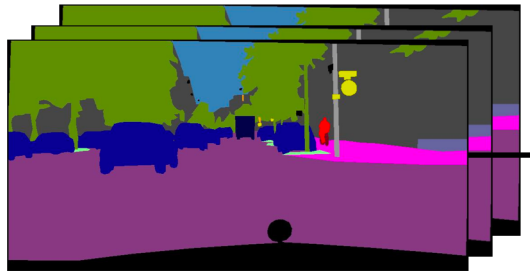


X

Discriminator

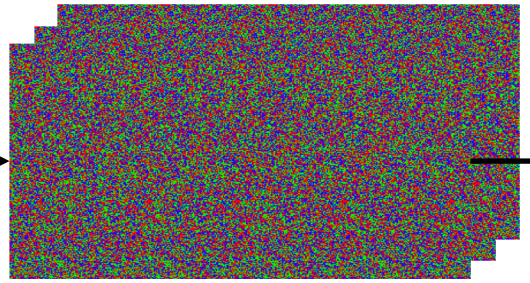
True

Introduction – Video Conditional GANs



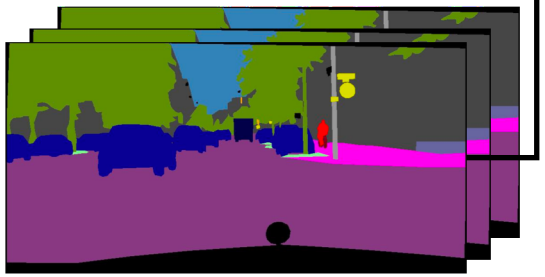
$\{X_i\}$

Generator



$\{G(X_i)\}$

Discriminator → False

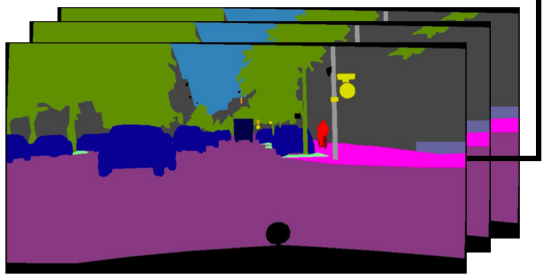


$\{X_i\}$

$p(\{G(X_i)\}, \{X_i\}) \rightarrow p(\{Y_i\}, \{X_i\})$

$\{Y_i\}$

Discriminator → True



$\{X_i\}$

Introduction – Related Video Synthesis Problems

- Future Video Prediction

  Finn et. al. 2016, Mathieu et. al. 2016, Lotter et. al. 2017, Xue et. al. 2016, Walker et. al. 2016, Denton et. al. 2017, Liang et. al. 2017, Villegas et. al. 2017 ….

- Unconditional Video Synthesis

  Vondrick et. al. 2016, Tulyakov et. al. 2017, Saito et. al. 2017, …

- Video super-resolution

  Shechtman et. al. 2005, Shi et. al. 2016, …

- Video inpainting

  Wexler et. al. 2004, 2007, …

- Motion Retargeting

  CHAN et. al. 2018, …

Outline

- Introduction
- **Method**
- Results
- Next Frame Prediction
- Conclusion

Method

	pix2pix -> pix2pixHD	pix2pixHD -> vid2vid
Generator	Coarse-to-fine image generator	Sequential generator Warping + Hallucination
Discriminator	Multi-scale conditional image discriminator	Multi-scale flow-conditioned spatial-temporal discriminator
Learning	GAN feature matching loss	Spatio-temporally Progressive Training

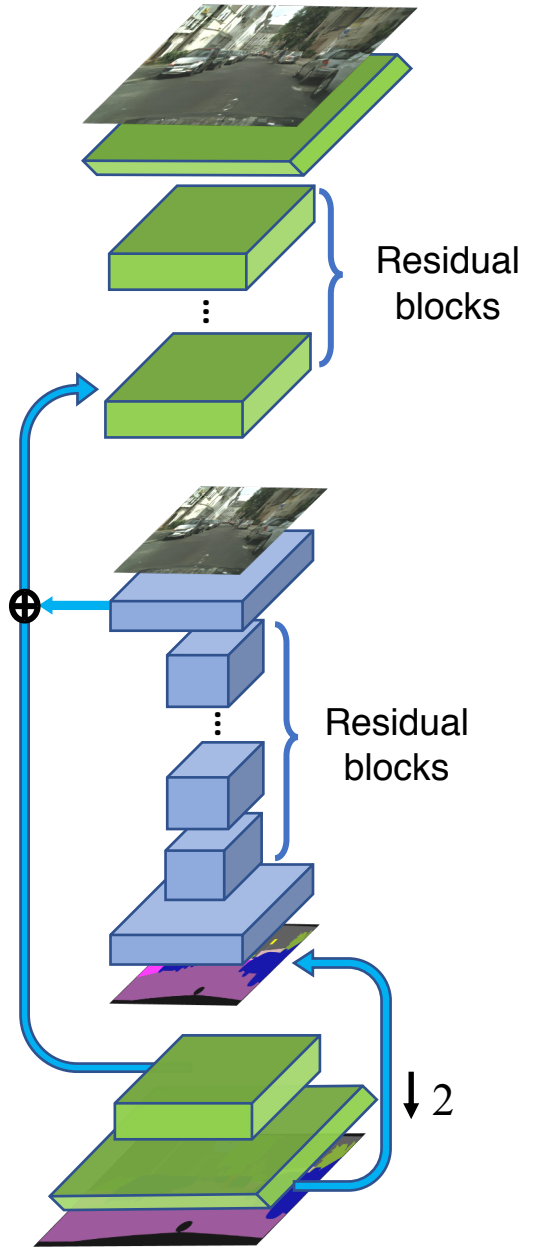
Method – Review pix2pixHD

Coarse-to-fine Generator

Multi-scale Discriminators

Robust Objective

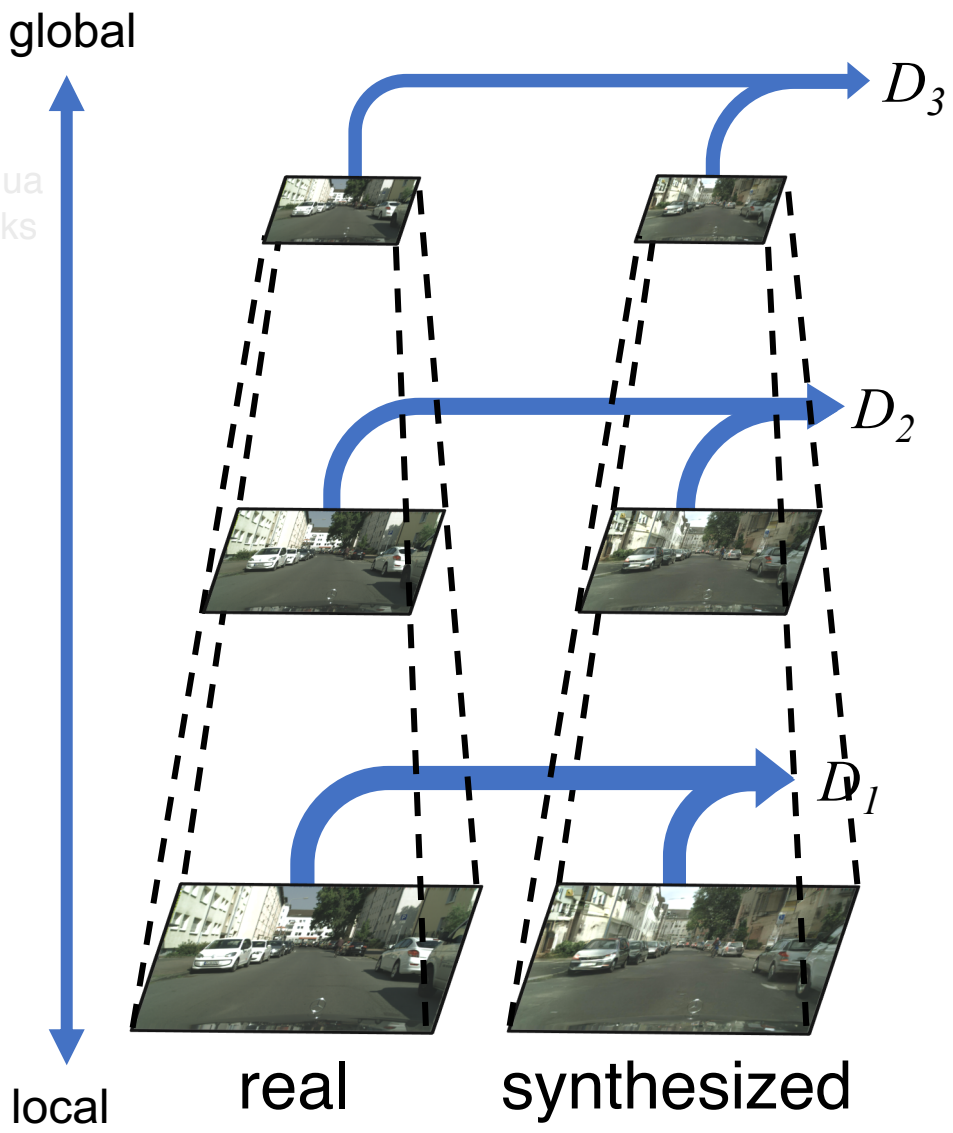
Coarse-to-fine Generator



Multi-scale Discriminators

Robust Objective

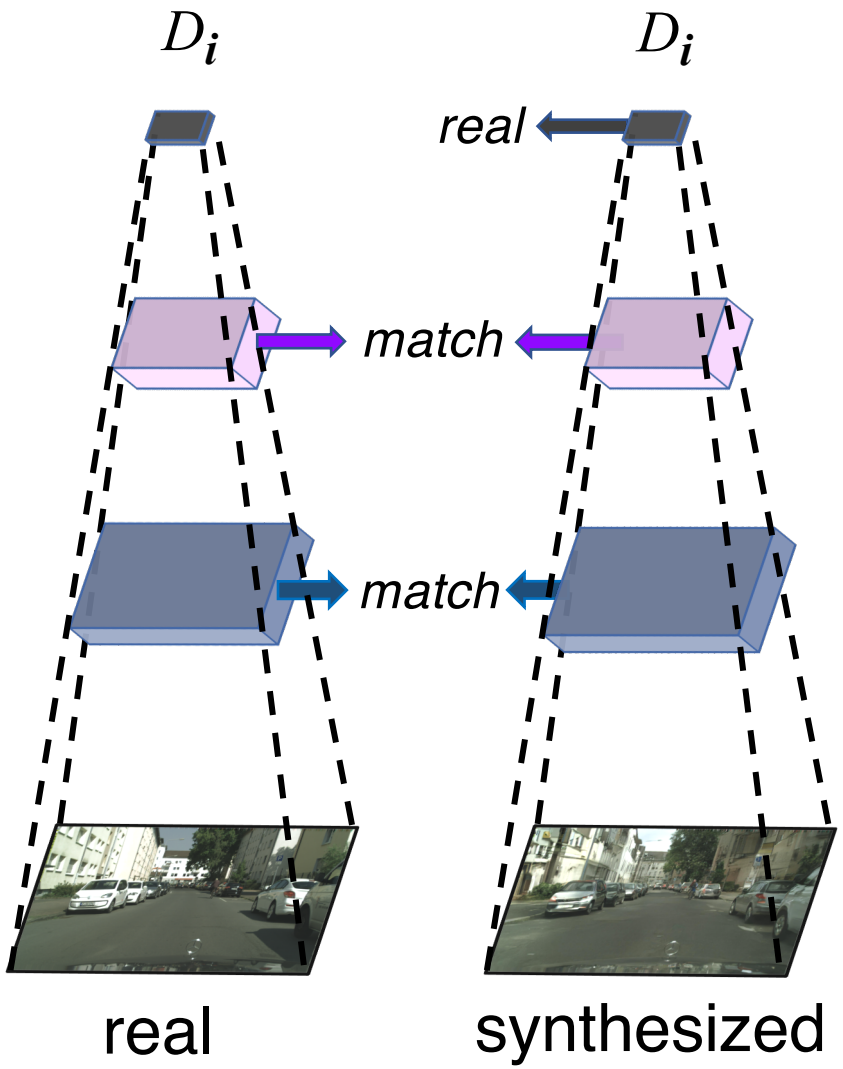
Multi-scale Discriminators



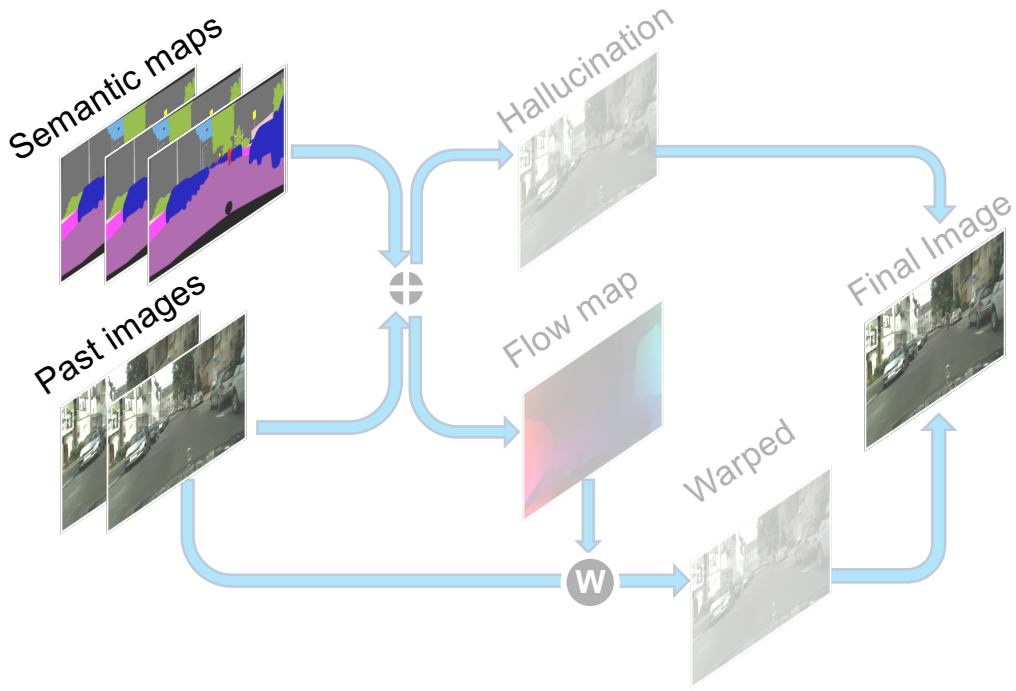
*Similar ideas in Durugkar et al. 2016, Iizuka et al. 2017, Zhang et al. 2017

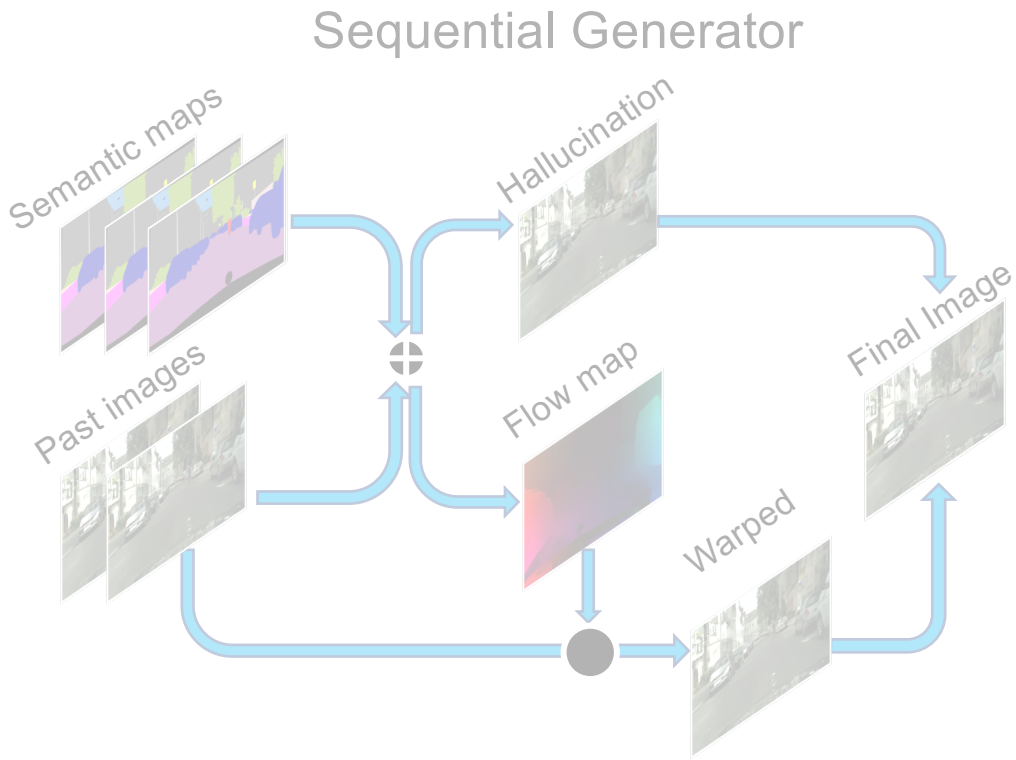
Coarse-to-fine Generator Multi-scale Discriminators

Robust Objective

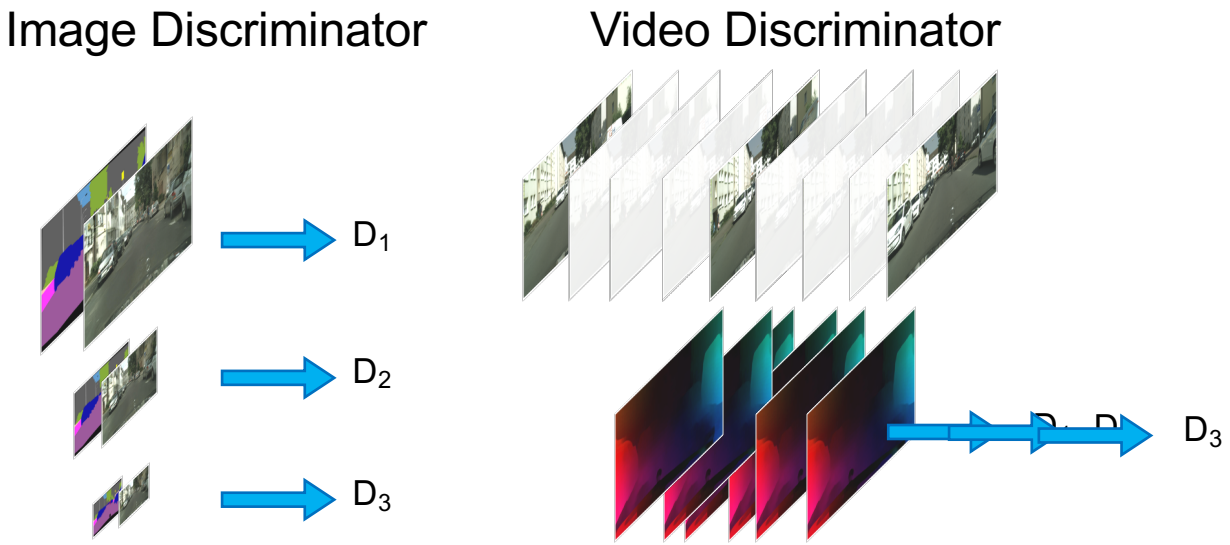


Method – New Things in vid2vid

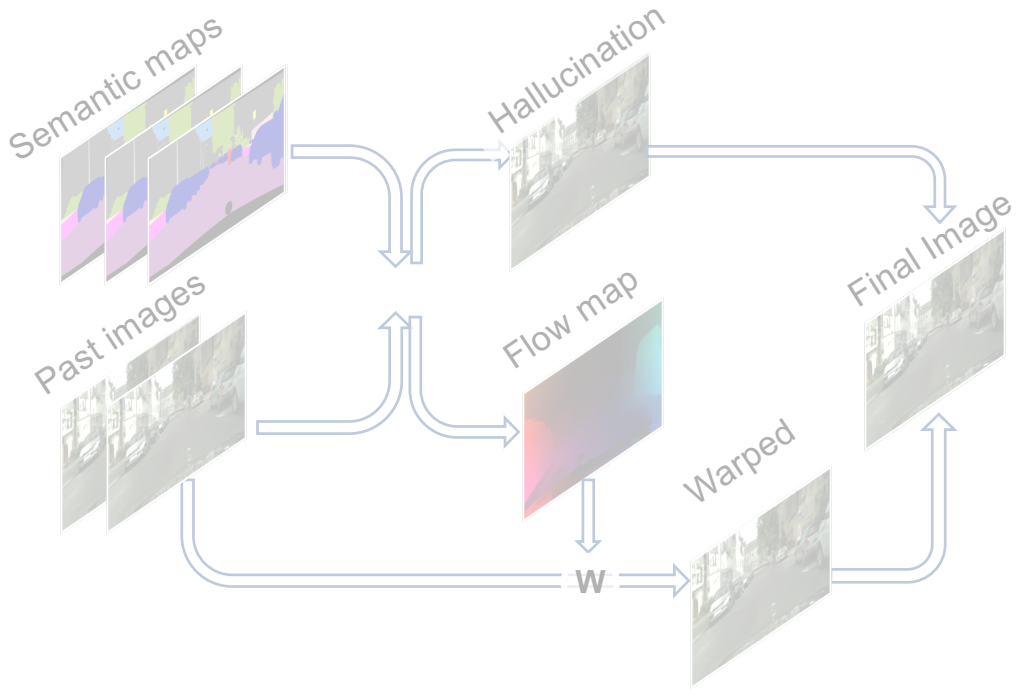

Semantic maps
Past images
Hallucination
Flow map
Warped
Final Image
W



Multi-scale Discriminators



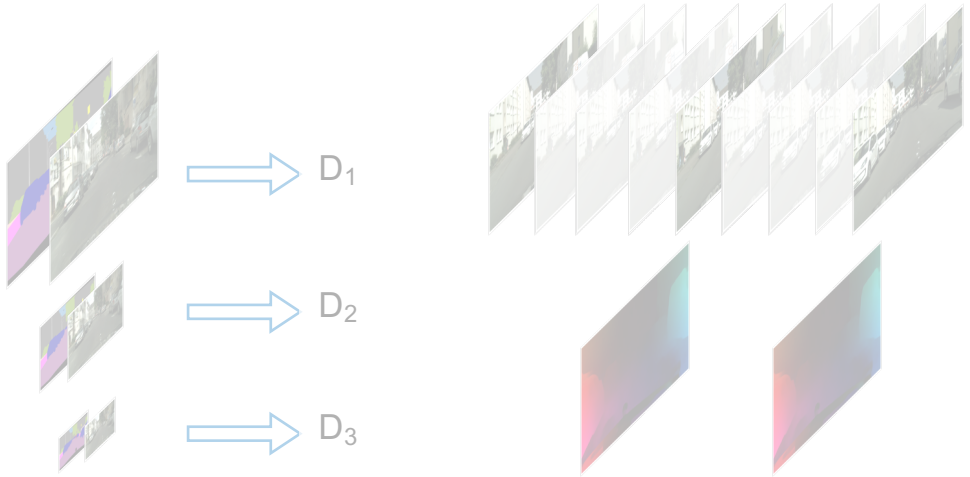
Sequential Generator



Multi-scale Discriminators

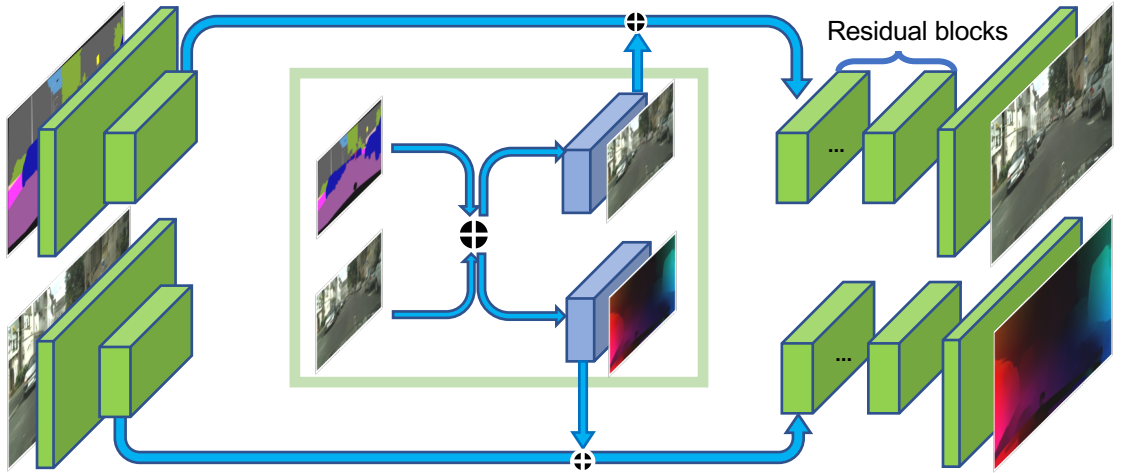
Image Discriminator

Video Discriminator

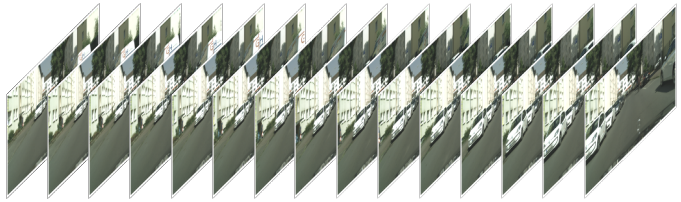


Spatio-temporally Progressive Training

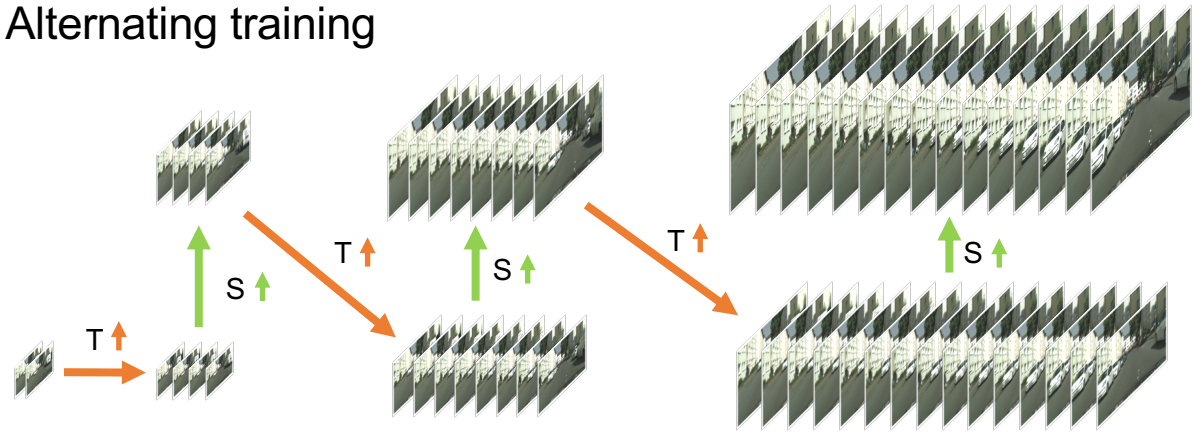
Spatially progressive



Temporally progressive



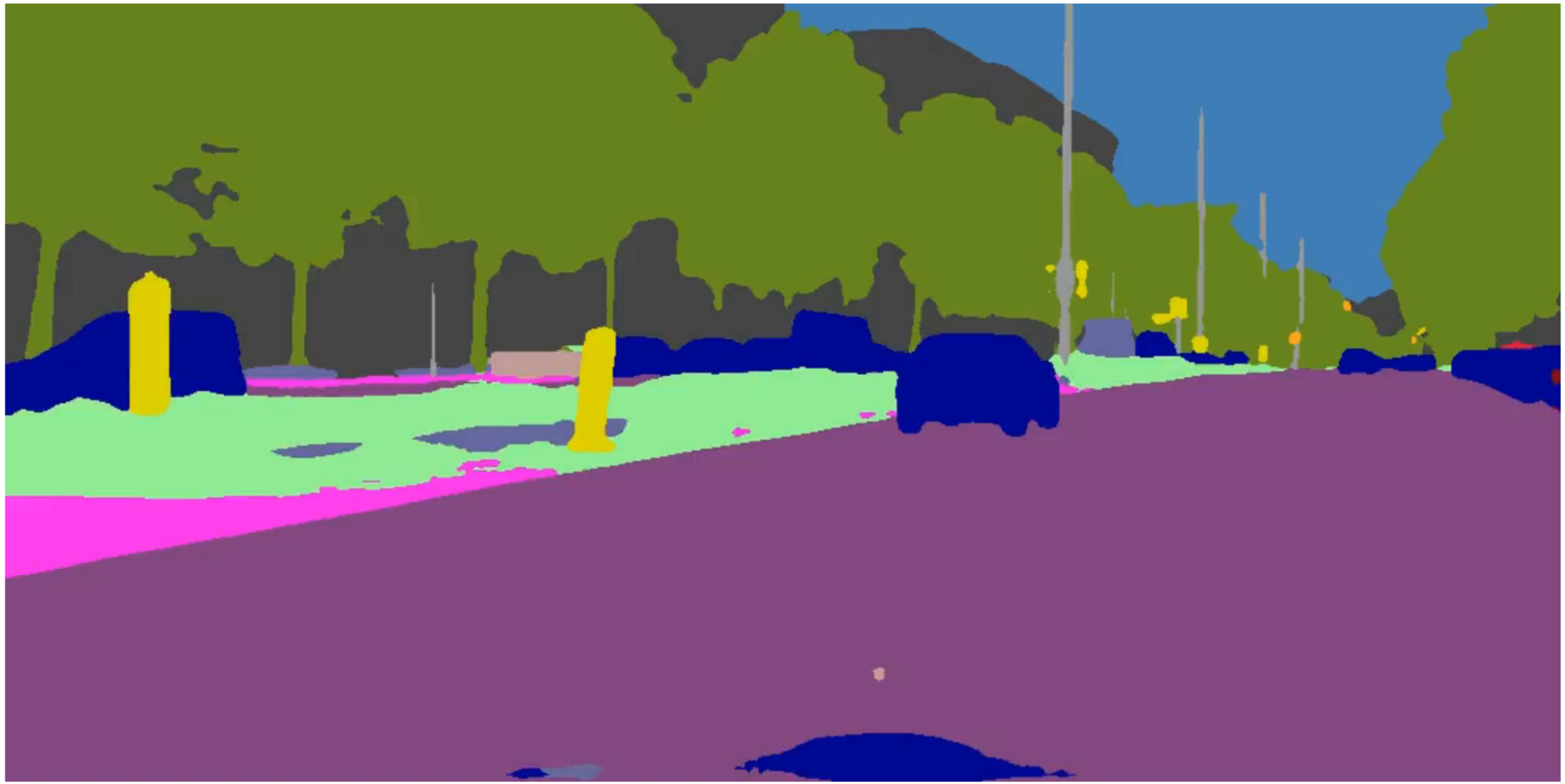
Alternating training



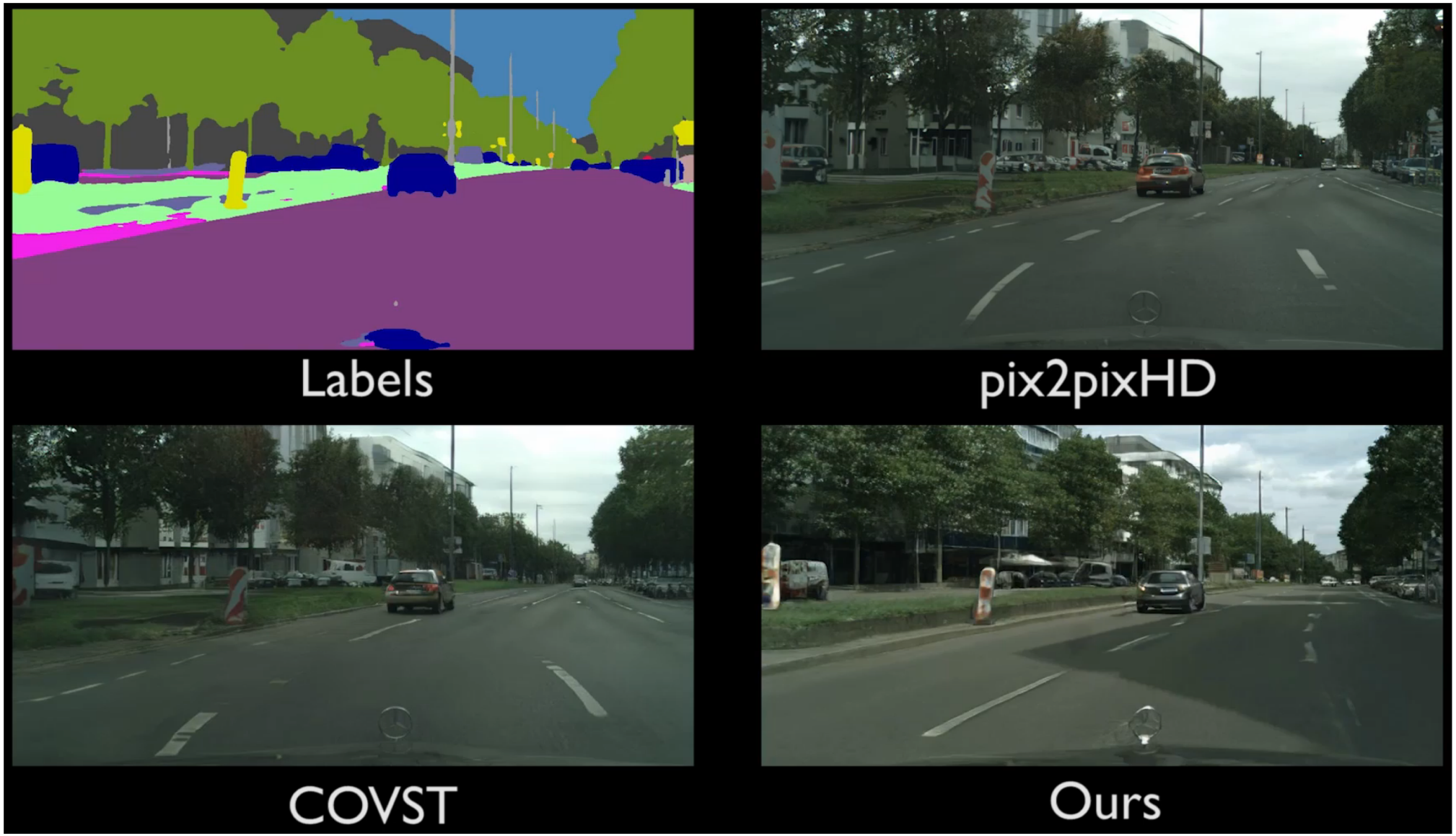
Outline

- Introduction
- Method
- **Results**
- Next Frame Prediction
- Conclusion

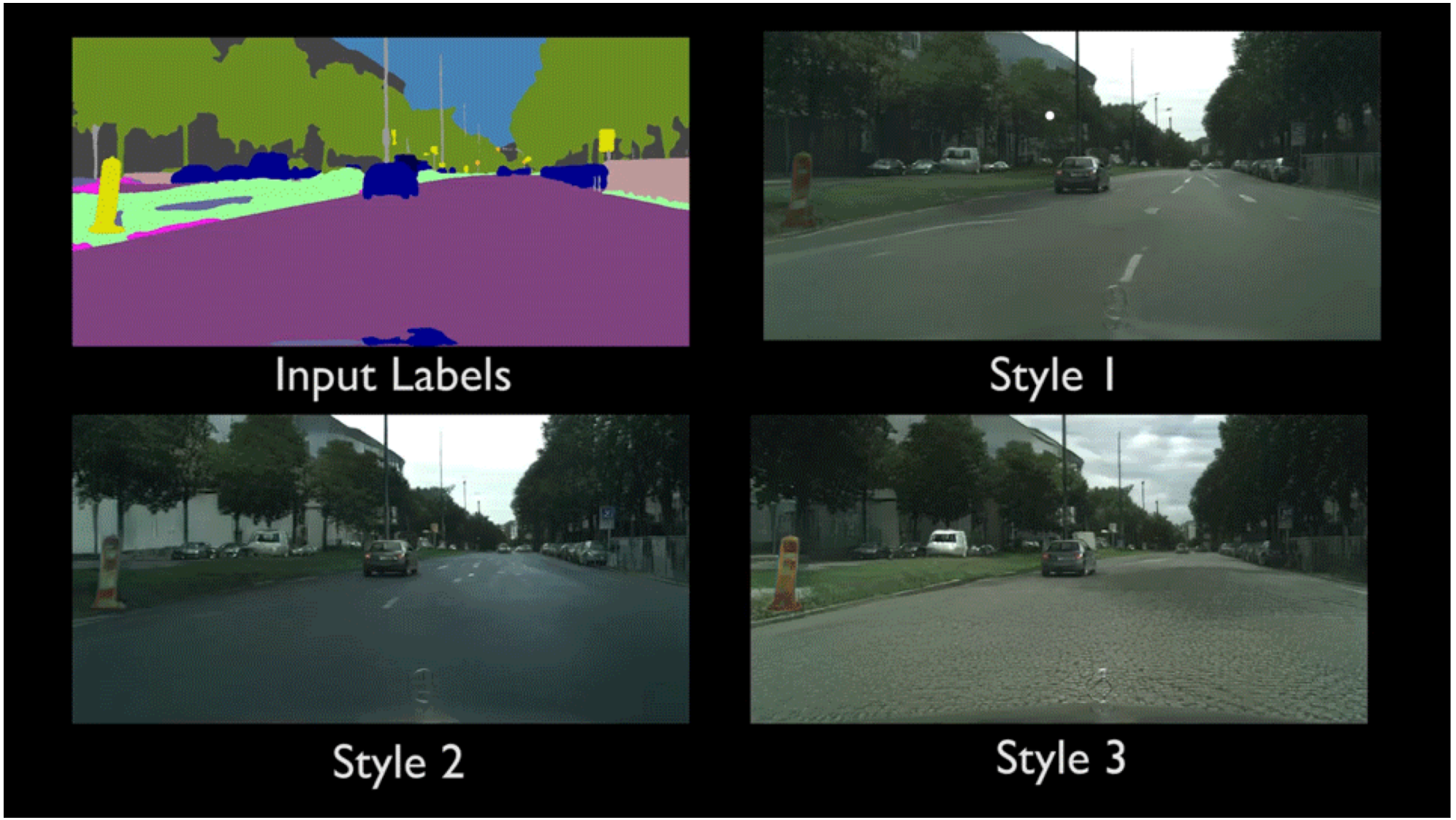
Results - Street view



Results - Street view



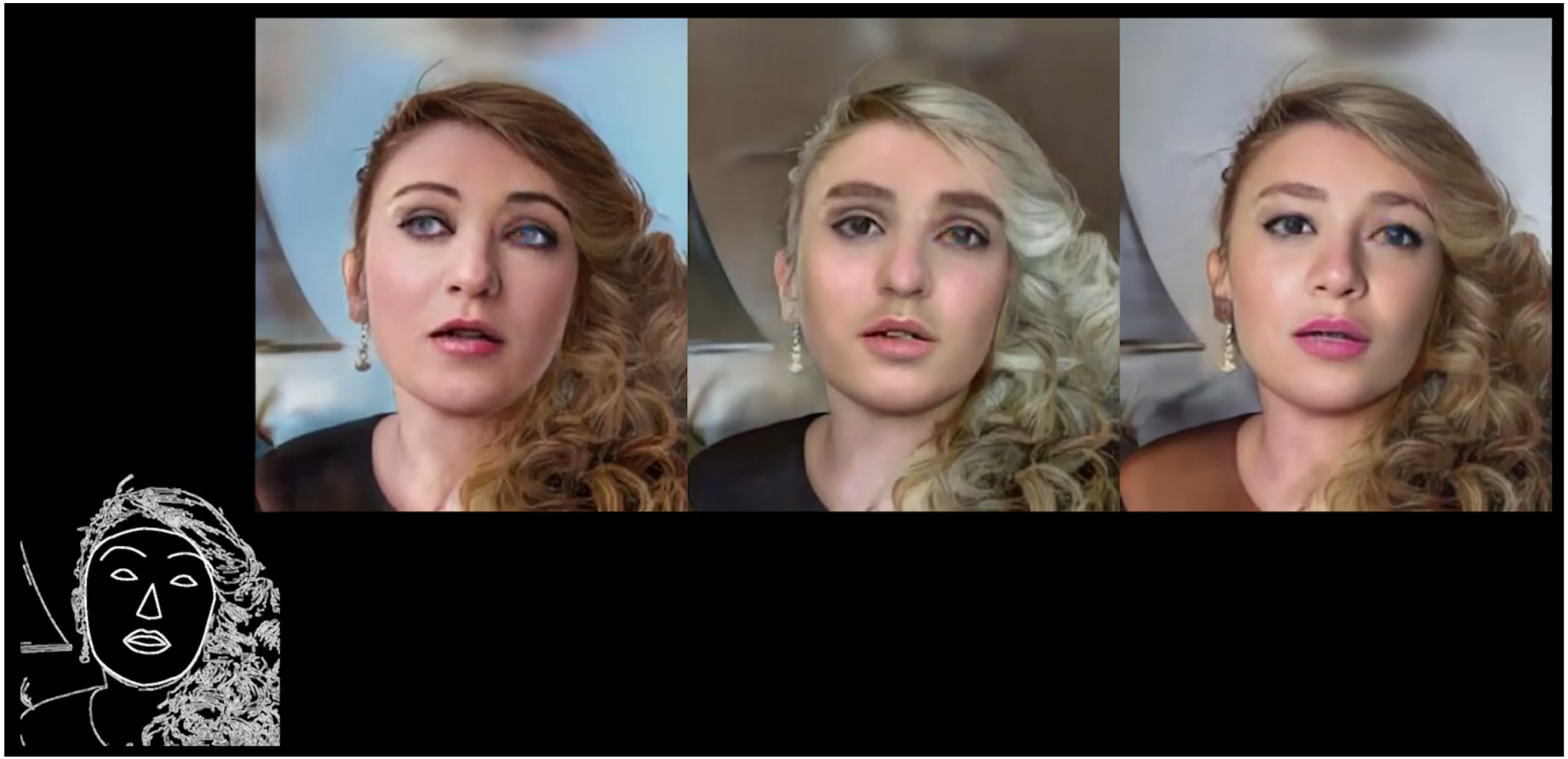
Results – Multimodal street scene



Results – Semantic Manipulation

Results - edge-to-face

Results - multimodal edge-to-face

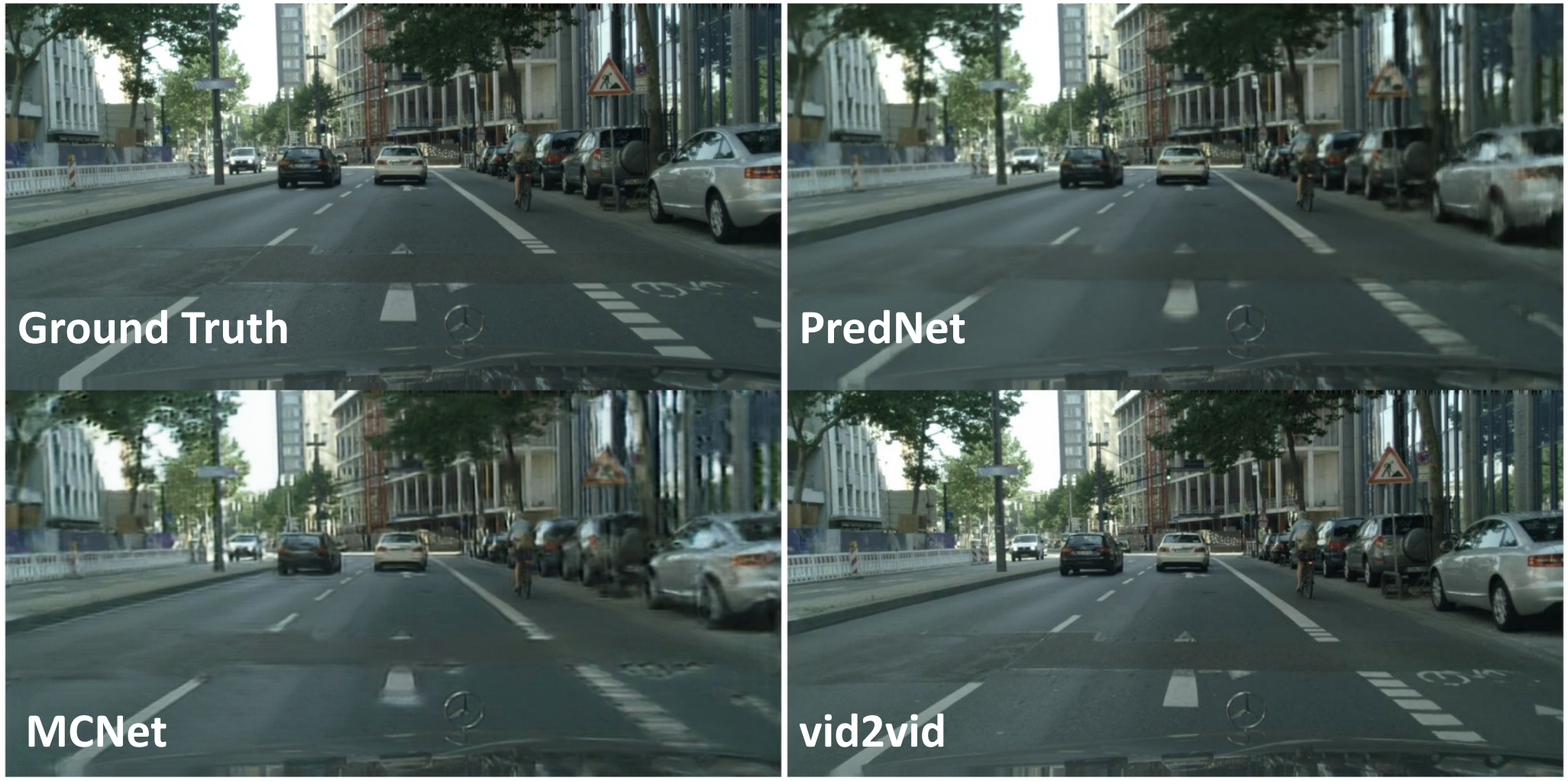


Results - pose-to-body

Outline

- Introduction
- Method
- Results
- **Next Frame Prediction**
- Conclusion

Next Frame Prediction

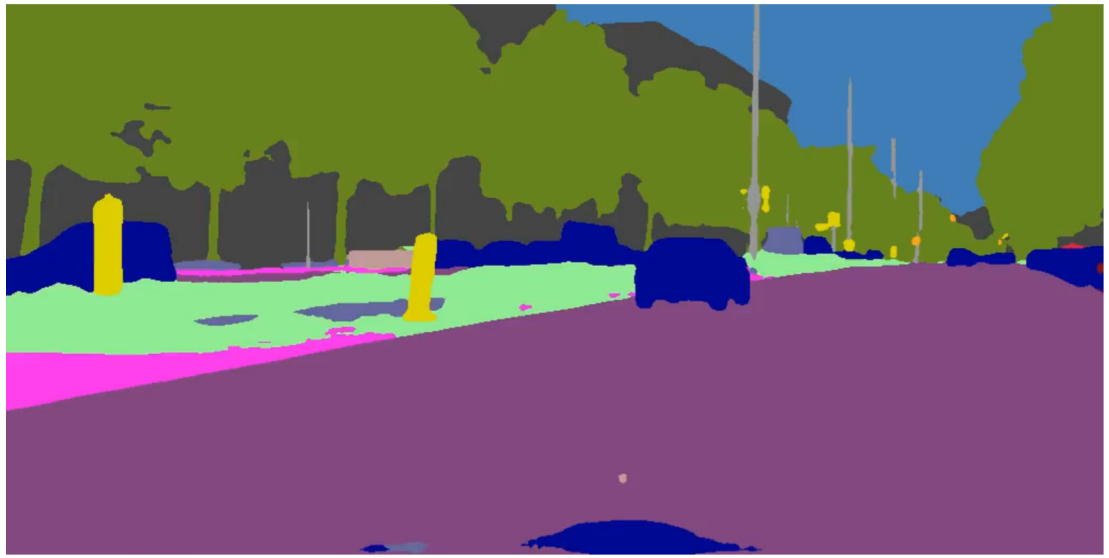


Outline

- Introduction
- Method
- Results
- Next Frame Prediction
- **Conclusion**

Conclusion – Failure Cases

- Turning cars
- Diversity of object appearances drop from image to video synthesis
- Does not guarantee long-term consistency.



Conclusion

- We present a general video-to-video synthesis framework.
  - Segmentation to image
  - Edges to image
  - Pose to image
- New Things
  - A sequential generator that synthesizes current frame based on warping and hallucination
  - A flow-conditioned temporal discriminator
  - Spatial-temporal progressive training

Thank you!

Paper: https://arxiv.org/abs/1808.06601 (Will be presented in NIPS)
Code: https://github.com/NVIDIA/vid2vid