NVIDIA

SEMI-SUPERVISED DEEP LEARNING APPLICATIONS

Bryan Catanzaro, 19 March 2019

SUPERVISED LEARNING

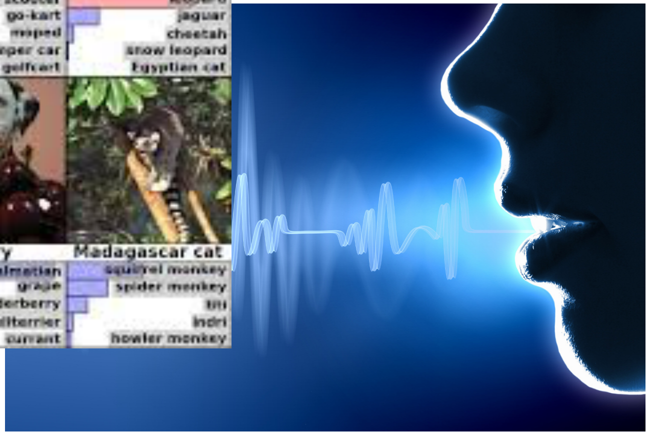
Mappings from X -> Y

Image classification

Speech recognition

Recommendation systems

Natural language understanding



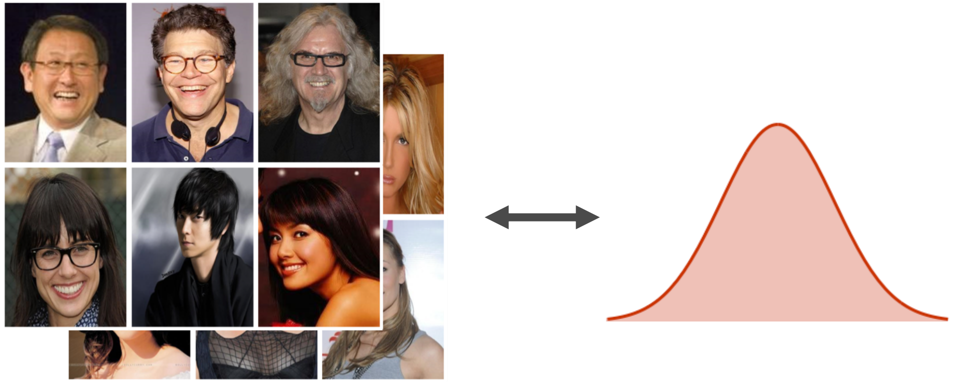
Works, but labeling data is slow and expensive

SEMI-SUPERVISED LEARNING

Learn data distributions from unlabeled data

Make use of a few labels to solve the problem

But what can you do with a data distribution
+ a few labels?



Semantic segmentation

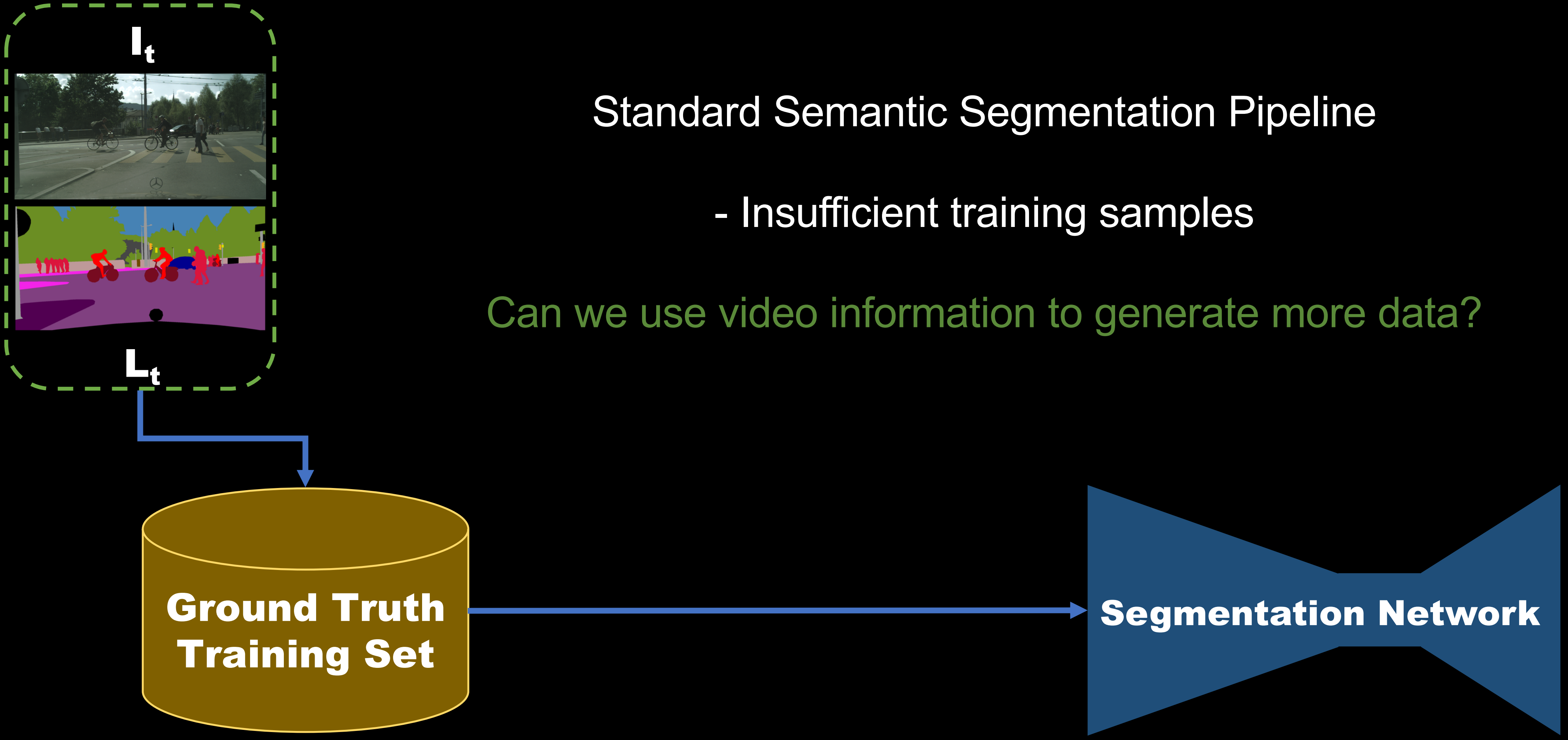
Image and video synthesis

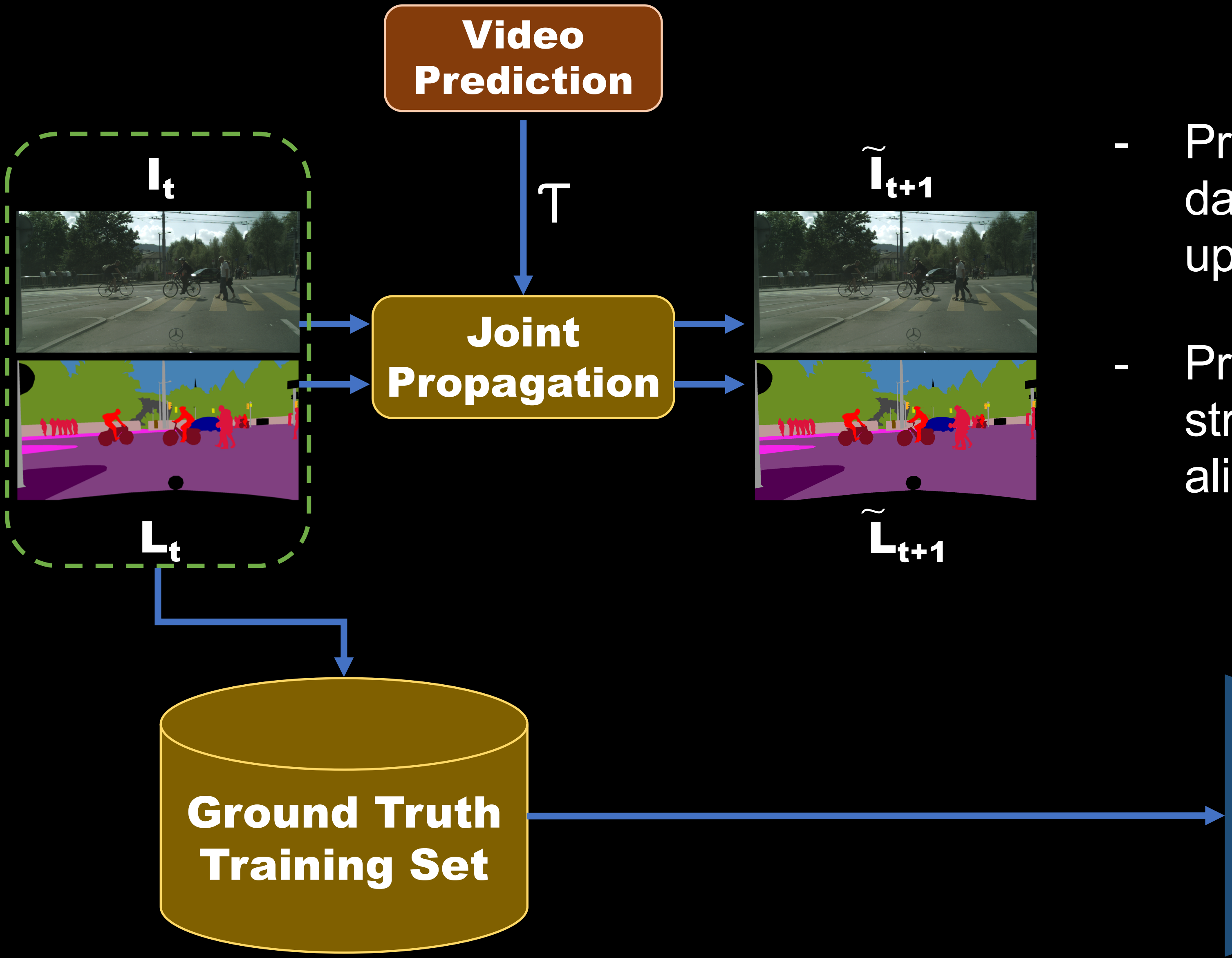
Text classification

Speech synthesis

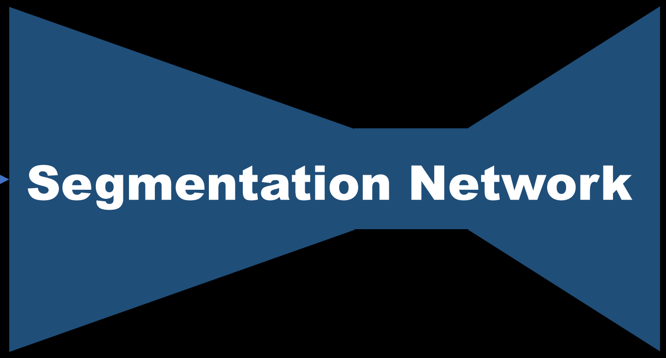
SEMANTIC SEGMENTATION

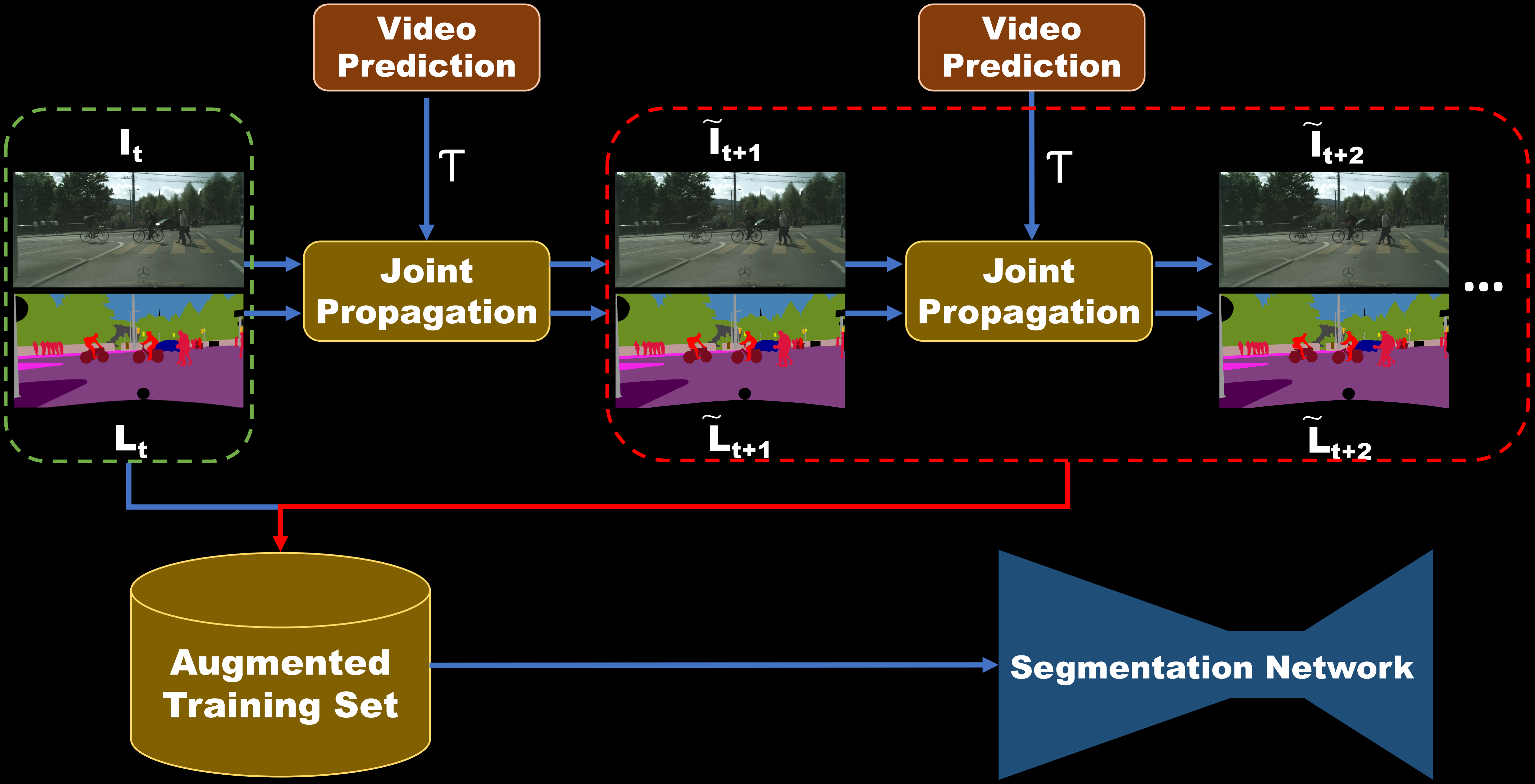
Yi Zhu, Karan Sapra et al., CVPR 2019

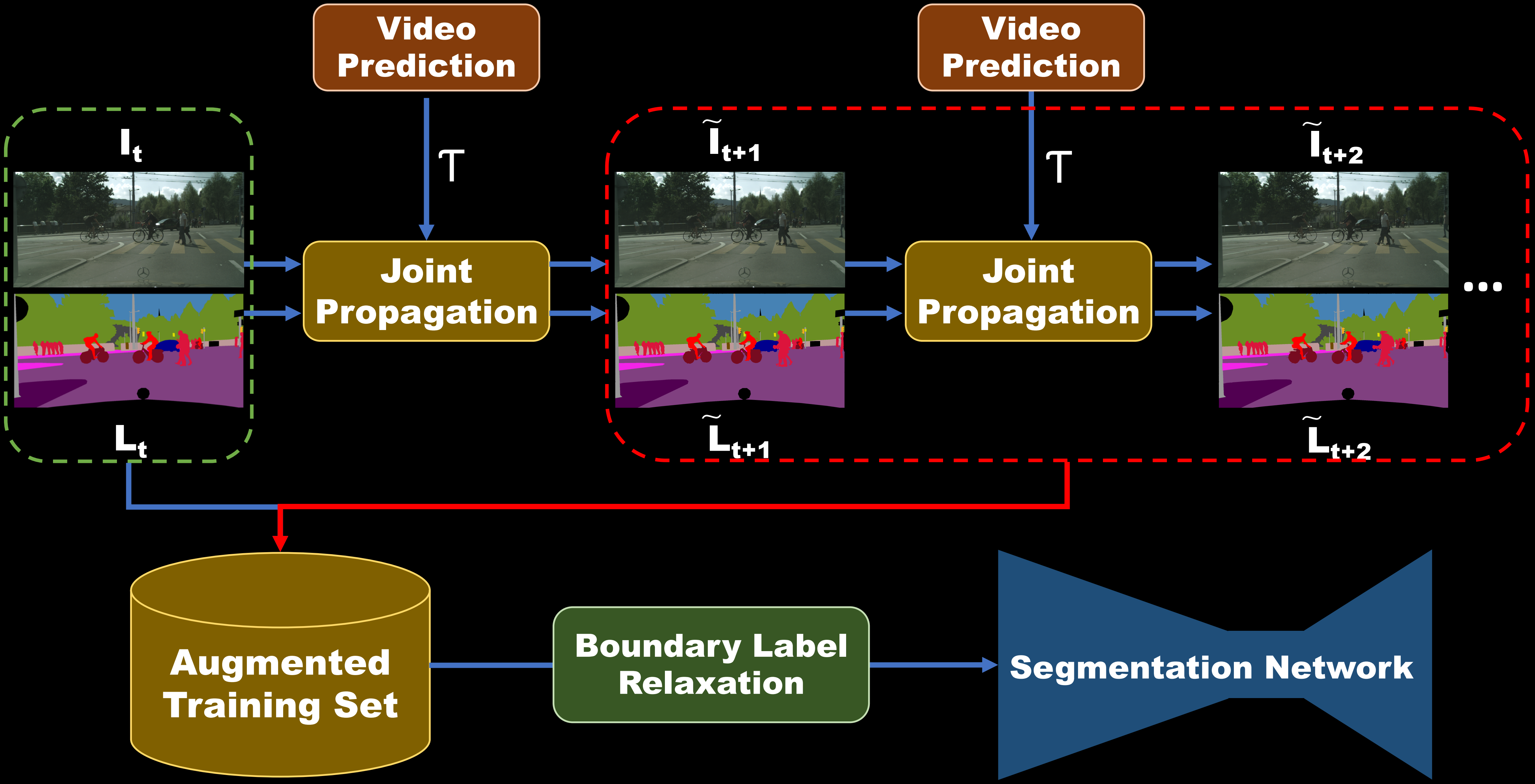

I_t
L_t
Standard Semantic Segmentation Pipeline
- Insufficient training samples
Can we use video information to generate more data?
Ground Truth Training Set
Segmentation Network



- Propose video prediction-based data synthesis method to scale up training sets
- Propose joint propagation strategy to alleviate mis-alignment problem




Video Prediction
Video Prediction
I_t
$\mathcal{T}$
Joint Propagation
$\tilde{I}_{t+1}$
$\mathcal{T}$
Joint Propagation
$\tilde{I}_{t+2}$
...
L_t
$\tilde{L}_{t+1}$
$\tilde{L}_{t+2}$
Augmented Training Set
Segmentation Network

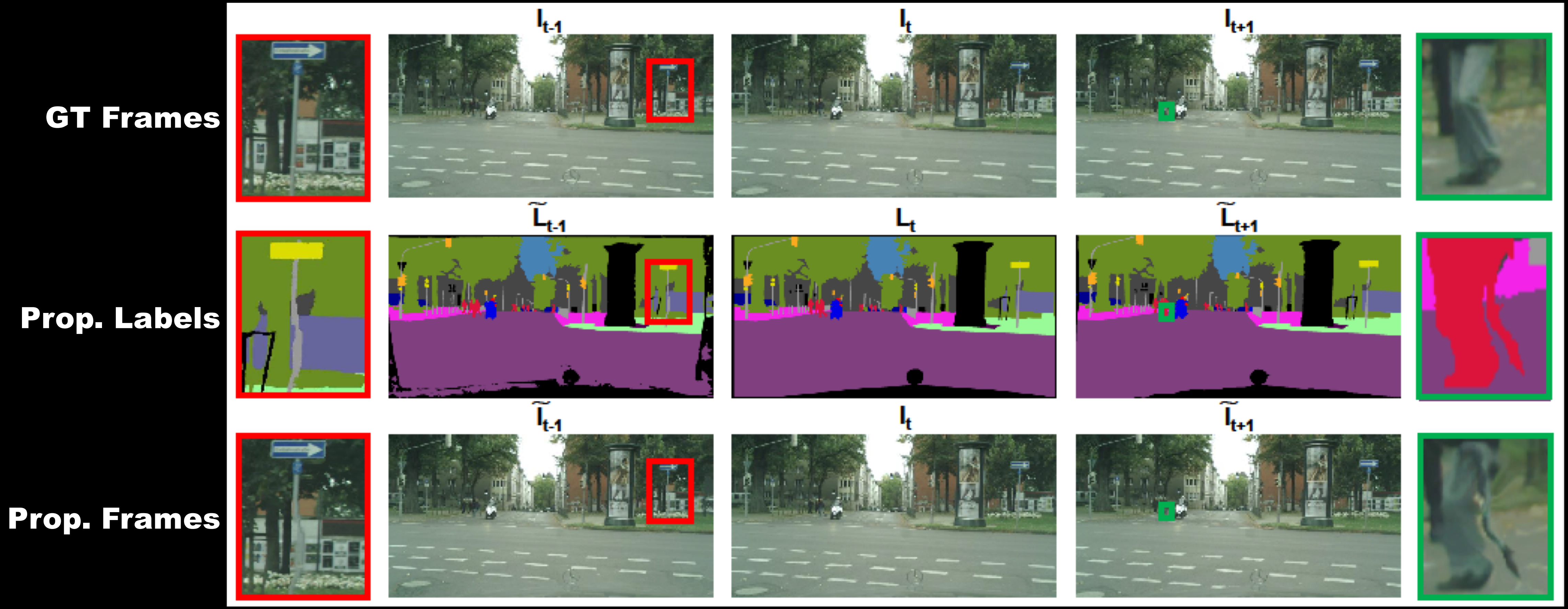


- Propose label relaxation to mitigate label noise during model training

Label Propagation

Mis-alignment problem between GT frames and propagated labels
e.g., street pole (red box) and person leg (green box).

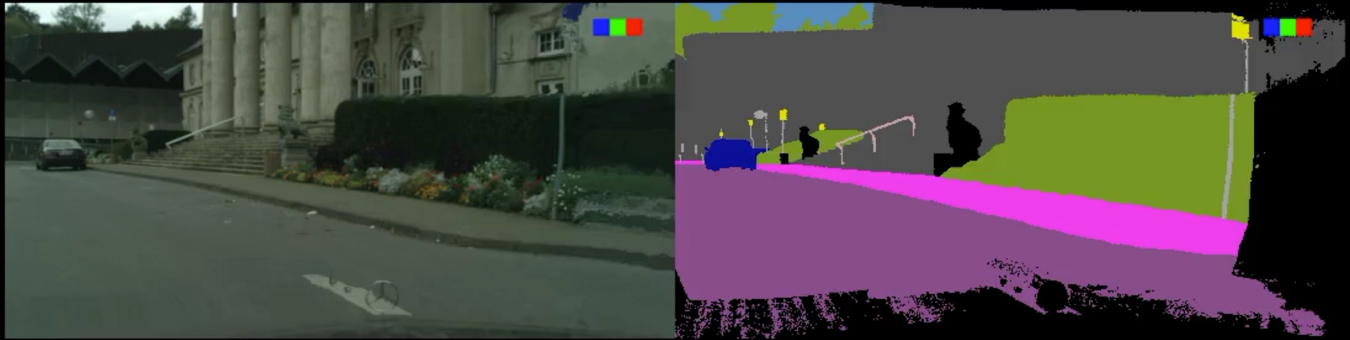
Joint Propagation



Higher degree of alignment between propagated frames and propagated labels.

Cityscapes Synthesized Training Samples

Propagation [t-5, t+5]



Joint Propagation using Video Prediction

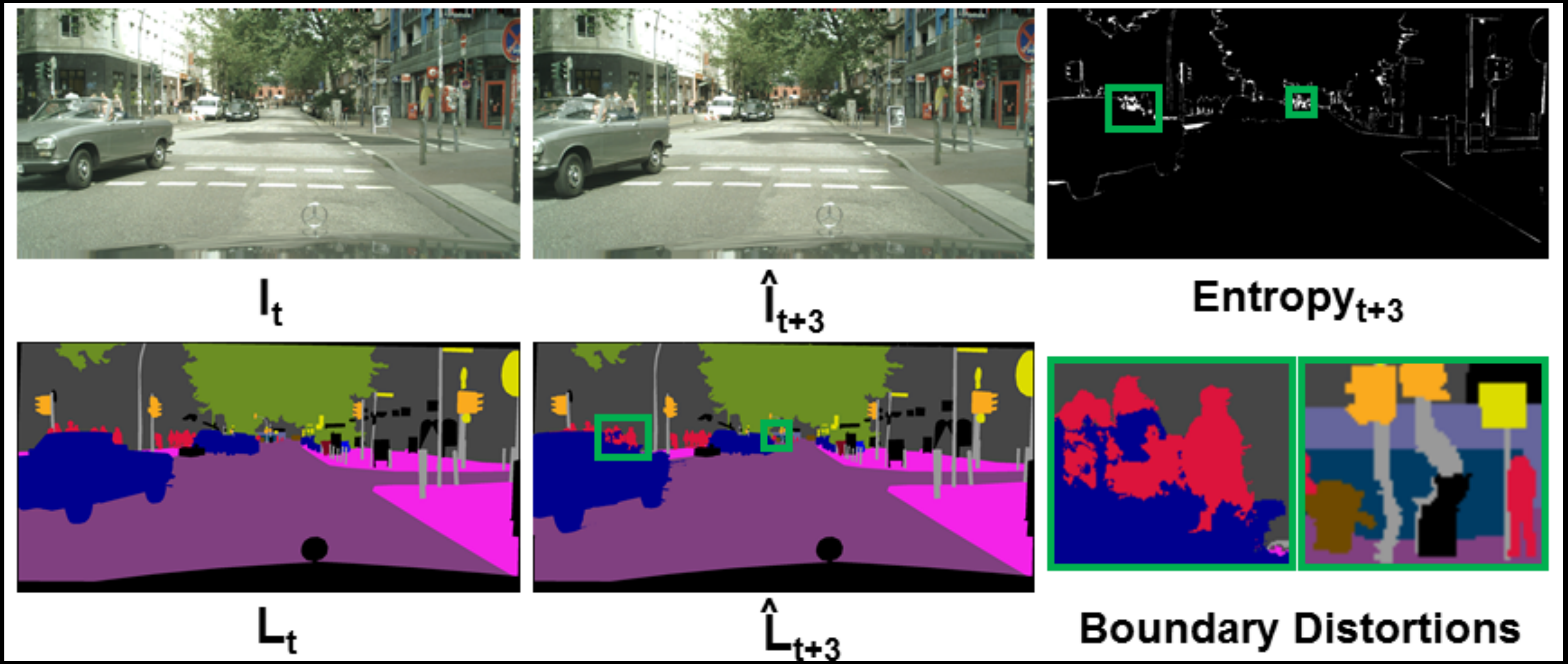
GT Frames

Joint Propagation using Video Reconstruction

synthesized samples

Joint Propagation using FlowNet2

Boundary Label Relaxation



Higher entropy
- ambiguous labeling
- propagation artifacts

We propose a modification to class label space that allows us to predict multiple classes at a boundary pixel

faster convergence, better generalization

Table 3: Per-class mIoU results on Cityscapes. Top: our ablation improvements on the validation set. Bottom: comparison with top-performing models on the test set.

Method	split	road	swalk	build.	wall	fence	pole	tlight	tsign	veg.	terrain	sky	person	rider	car	truck	bus	train	mcycle	bicycle	mIoU
Baseline	val	98.4	86.5	93.0	57.4	65.5	66.7	70.6	78.9	92.7	65.0	95.3	80.8	60.9	95.3	87.9	91.0	84.3	65.8	76.2	79.5
+ VRec with JP	val	98.0	86.5	94.7	47.6	67.1	69.6	71.8	80.4	92.2	58.4	95.6	88.3	71.1	95.6	76.8	84.7	90.3	79.6	80.3	80.5
+ Label Relaxation	val	98.5	87.4	93.5	64.2	66.1	69.3	74.2	81.5	92.9	64.6	95.6	83.5	66.5	95.7	87.7	91.9	85.7	70.1	78.8	81.4
ResNet38 [38]	test	98.7	86.9	93.3	60.4	62.9	67.6	75.0	78.7	93.7	73.7	95.5	86.8	71.1	96.1	75.2	87.6	81.9	69.8	76.7	80.6
PSPNet [43]	test	98.7	86.9	93.5	58.4	63.7	67.7	76.1	80.5	93.6	72.2	95.3	86.8	71.9	96.2	77.7	91.5	83.6	70.8	77.5	81.2
InPlaceABN [10]	test	98.4	85.0	93.6	61.7	63.9	67.7	77.4	80.8	93.7	71.9	95.6	86.7	72.8	95.7	79.9	93.1	89.7	72.6	78.2	82.0
DeepLabV3+ [14]	test	98.7	87.0	93.9	59.5	63.7	71.4	78.2	82.2	94.0	73.0	95.8	88.0	73.0	96.4	78.0	90.9	83.9	73.8	78.9	82.1
DRN-CRL [45]	test	98.8	87.7	94.0	**65.1**	64.2	70.1	77.4	81.6	93.9	73.5	95.8	88.0	74.9	96.5	**80.8**	92.1	88.5	72.1	78.8	82.8
Ours	test	**98.8**	**87.8**	**94.2**	64.1	**65.0**	**72.4**	**79.0**	**82.8**	**94.2**	**74.0**	**96.1**	**88.2**	**75.4**	**96.5**	78.8	**94.0**	**91.6**	**73.8**	**79.0**	**83.5**

Table 4: Results on the CamVid test set. Pre-train indicates the source dataset on which the model is trained.

Method	Pre-train	Encoder	mIoU (%)
SegNet [3]	ImageNet	VGG16	60.1
RTA [19]	ImageNet	VGG16	62.5
Dilate8 [42]	ImageNet	Dilate	65.3
BiSeNet [41]	ImageNet	ResNet18	68.7
PSPNet [43]	ImageNet	ResNet50	69.1
DenseDecoder [6]	ImageNet	ResNeXt101	70.9
VideoGCRF [11]	Cityscapes	ResNet101	75.2
Ours (baseline)	Cityscapes	WideResNet38	79.8
Ours	Cityscapes	WideResNet38	**81.7**

Table 5: Results on KITTI test set.

Method	IoU class	iIoU class	IoU category	iIoU category
APMoE_seg [23]	47.96	17.86	78.11	49.17
SegStereo [40]	59.10	28.00	81.31	60.26
AHiSS [30]	61.24	26.94	81.54	53.42
LDN2 [24]	63.51	28.31	85.34	59.07
MapillaryAI [10]	69.56	43.17	86.52	68.89
Ours	**72.83**	**48.68**	**88.99**	**75.26**

IMAGE AND VIDEO SYNTHESIS

https://github.com/NVIDIA/vid2vid

Goal: render graphics with generative models

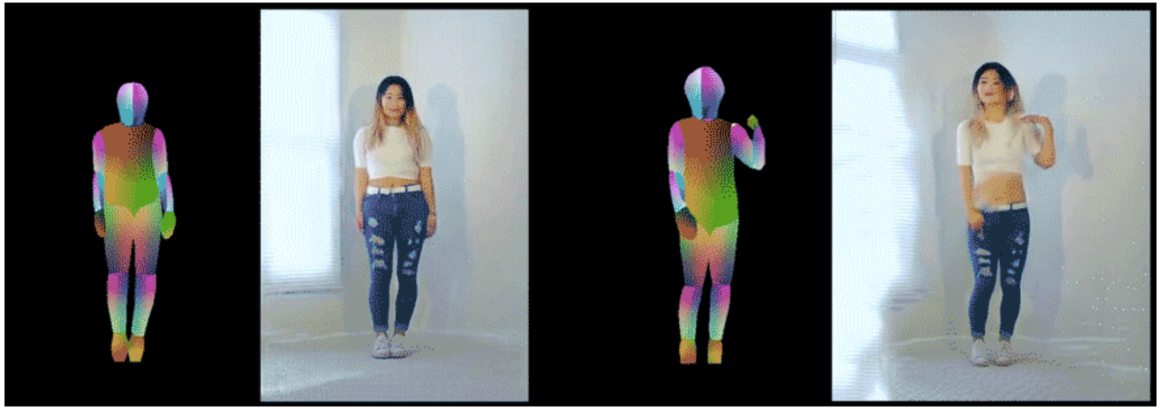
- We use a GAN

Condition on high level input

- Semantic map, edge map
- Easy to create and edit
- Provides control

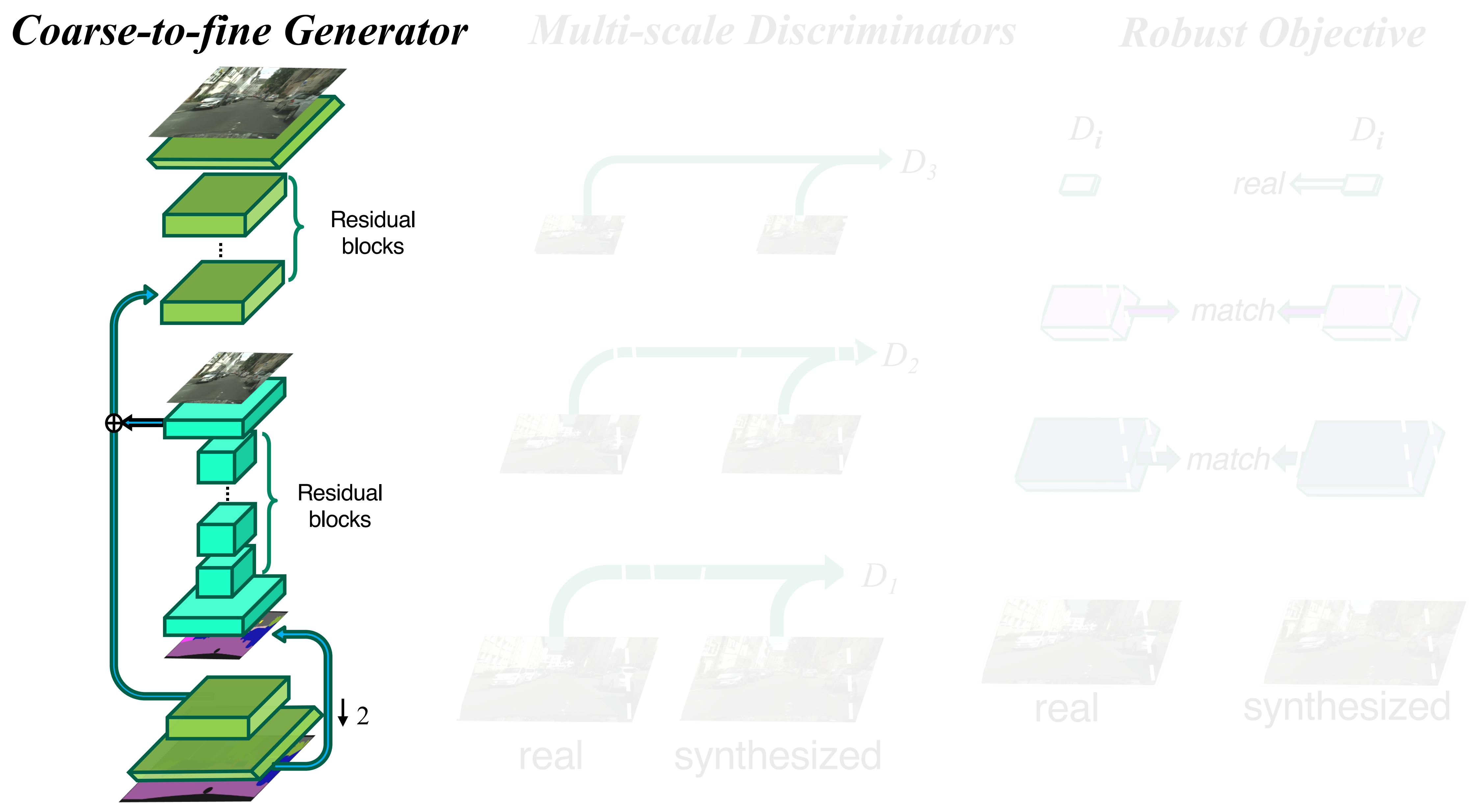
Render high resolution images

Create videos with temporal consistency



Input: pose map Output: rendered person

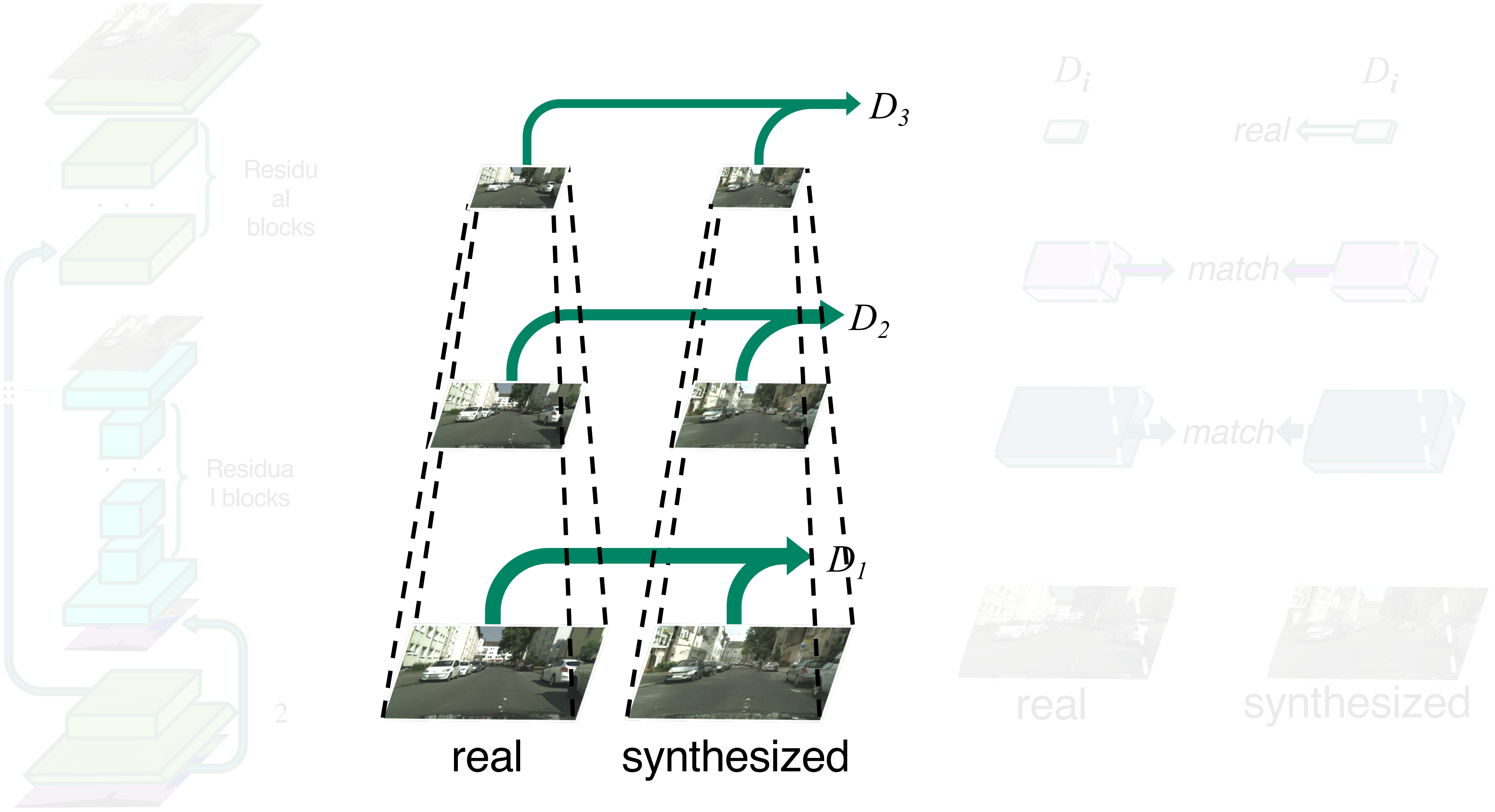
 Ting-Chun Wang et al., CVPR 2018, NeurIPS 2018

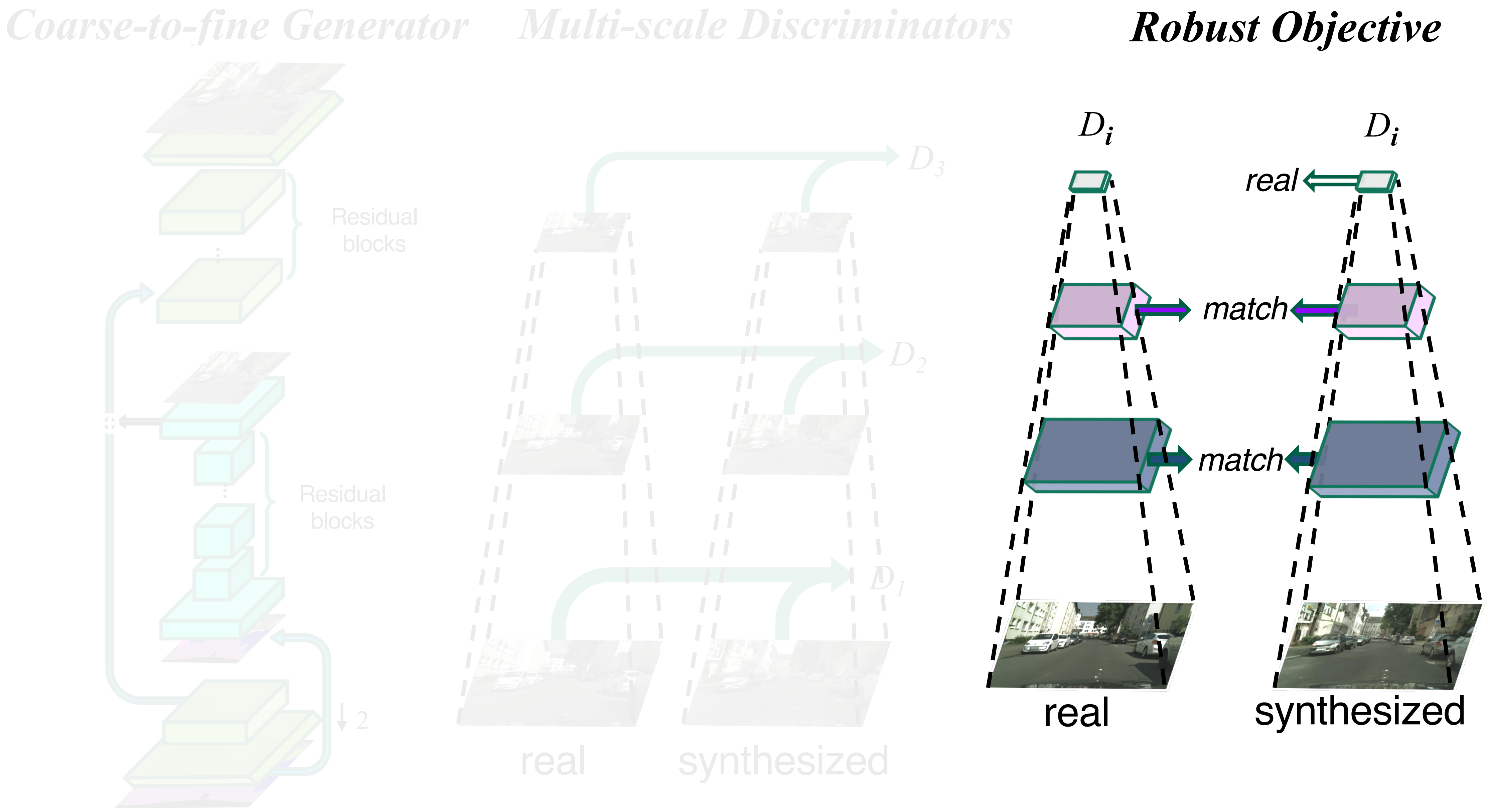

Coarse-to-fine Generator
Multi-scale Discriminators
Robust Objective
Residual blocks
Residual blocks
↓ 2
D3
D2
D1
real
synthesized
Di
Di
real
match
match
real
synthesized

Coarse-to-fine Generator

Multi-scale Discriminators

Robust Objective

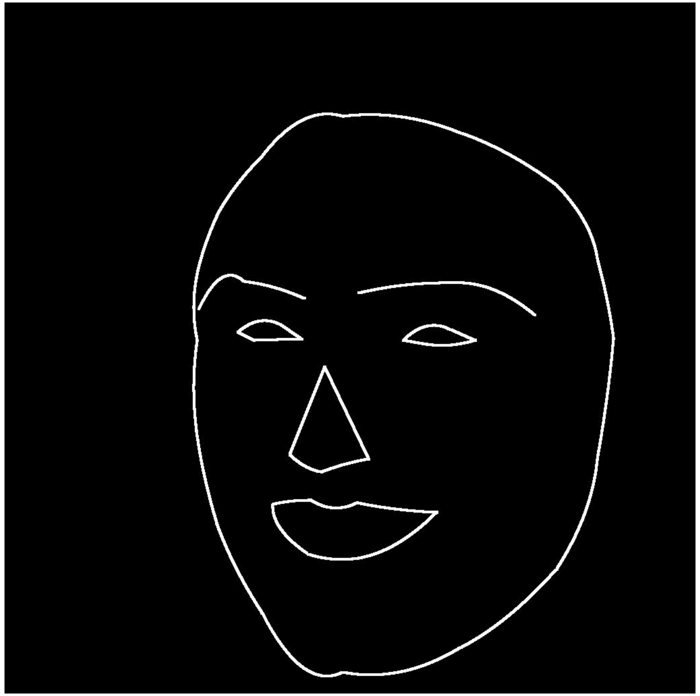




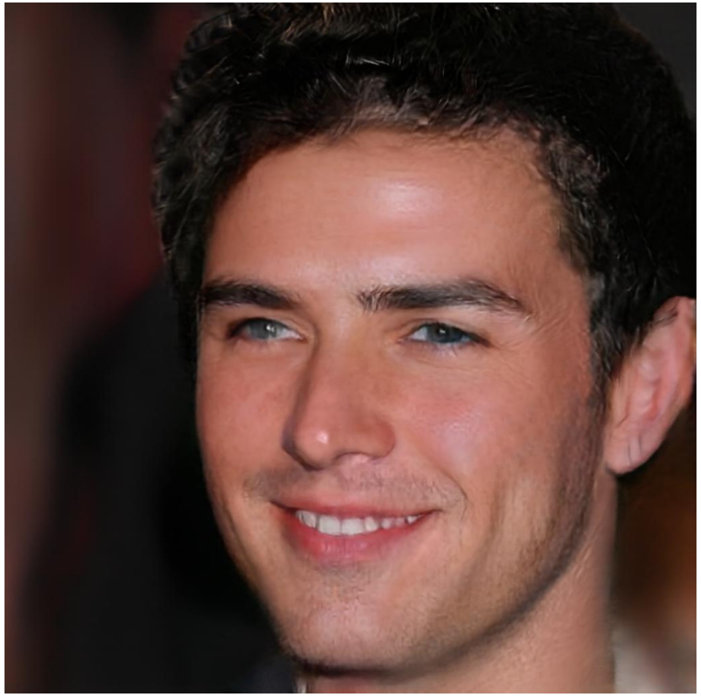
Coarse-to-fine Generator
Multi-scale Discriminators
Robust Objective
Residual blocks
Residual blocks
$\downarrow 2$
D_3
D_2
D_1
real
synthesized
D_i
D_i
real
match
match
real
synthesized

RESULTS

- CelebA-HQ



Edges

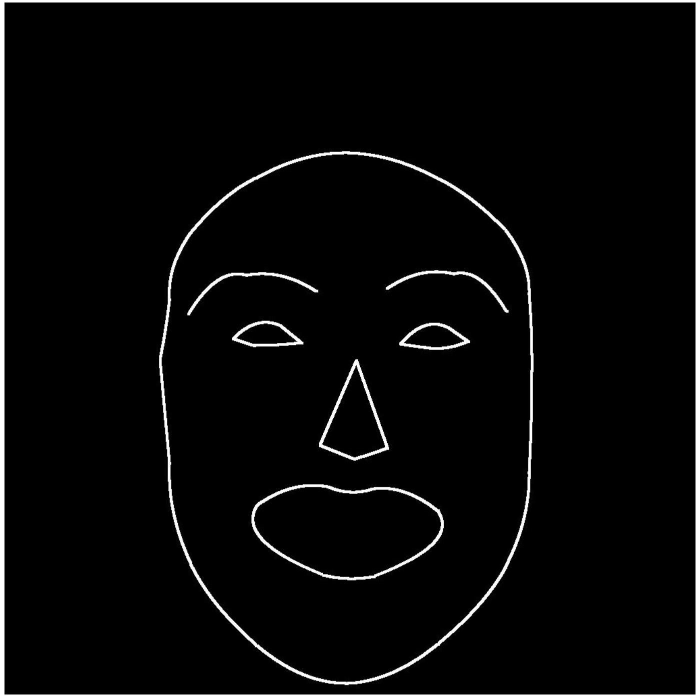


Synthesized

Ground truth

RESULTS

- CelebA-HQ



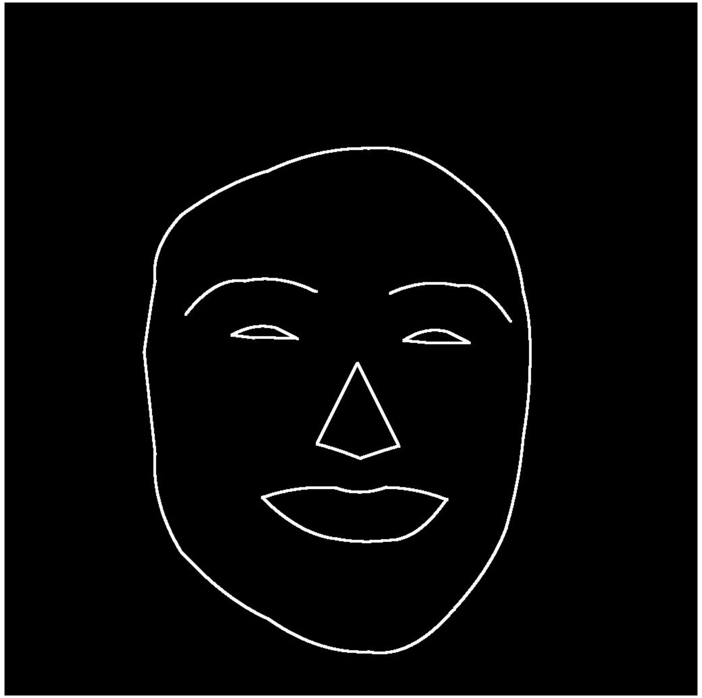
Edges

Synthesized

Ground truth

RESULTS

- CelebA-HQ



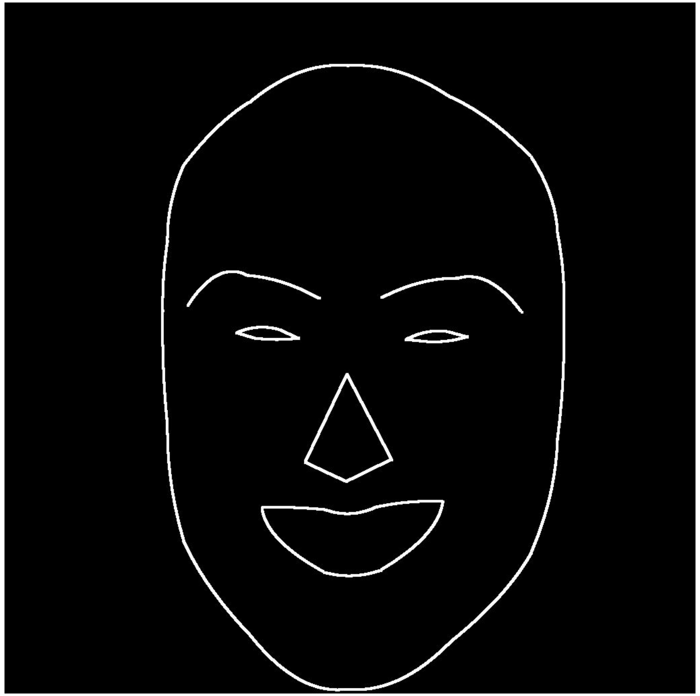
Edges

Synthesized

Ground truth

RESULTS

- CelebA-HQ

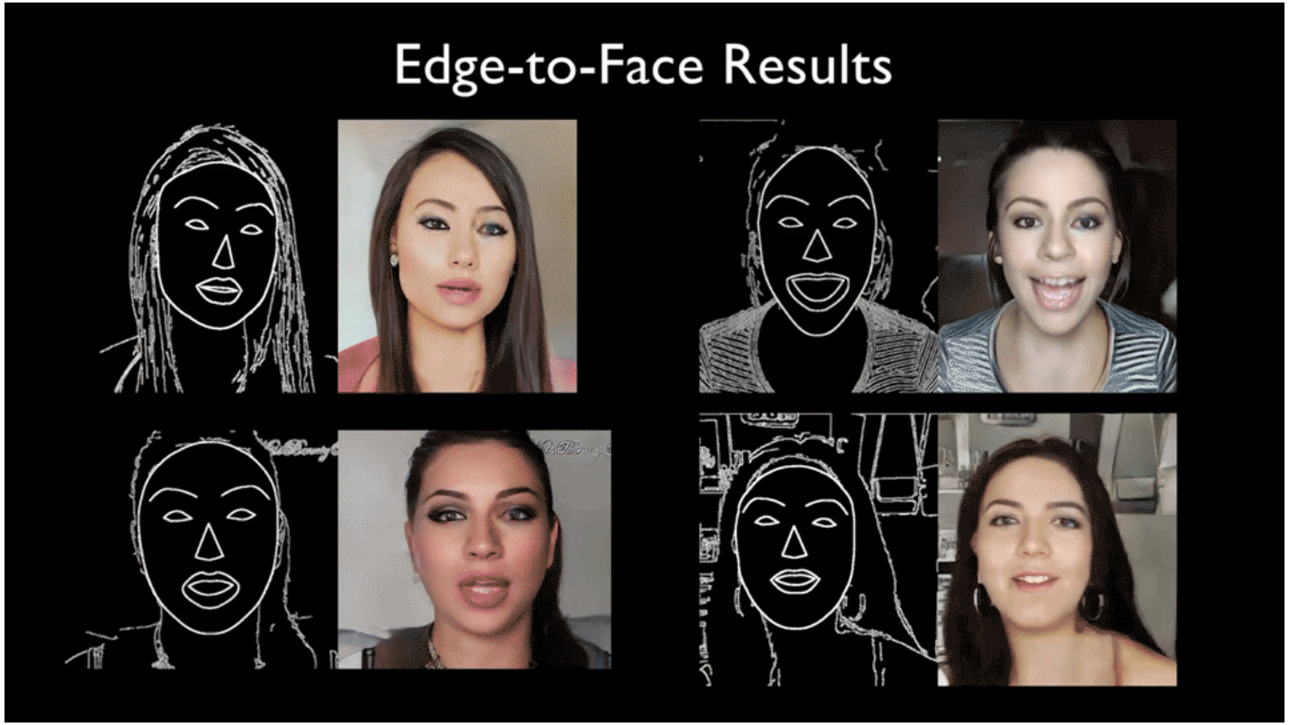


Edges

Synthesized

Ground truth

VIDEO SYNTHESIS

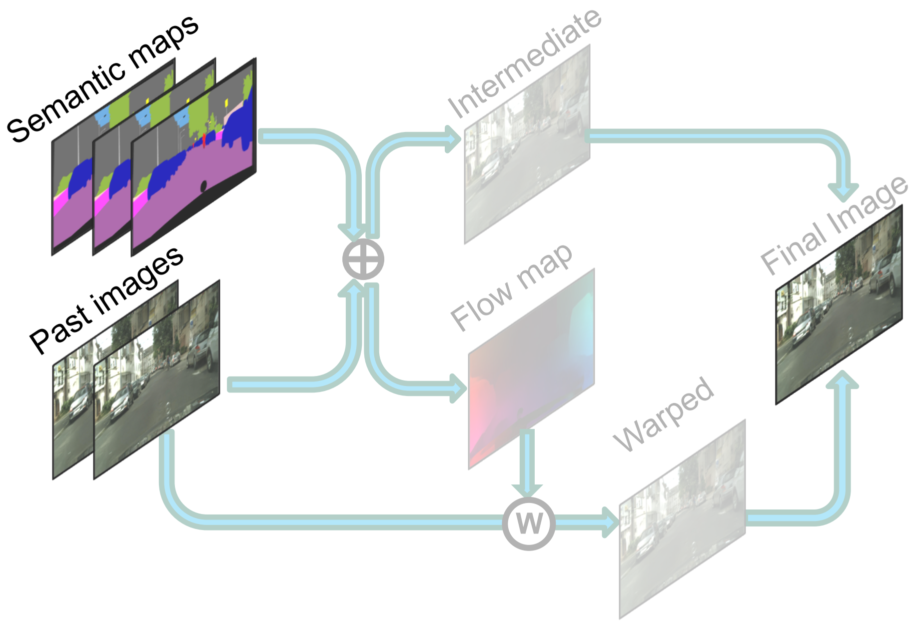


Input: edge maps

Output: neural network rendering

OUR METHOD

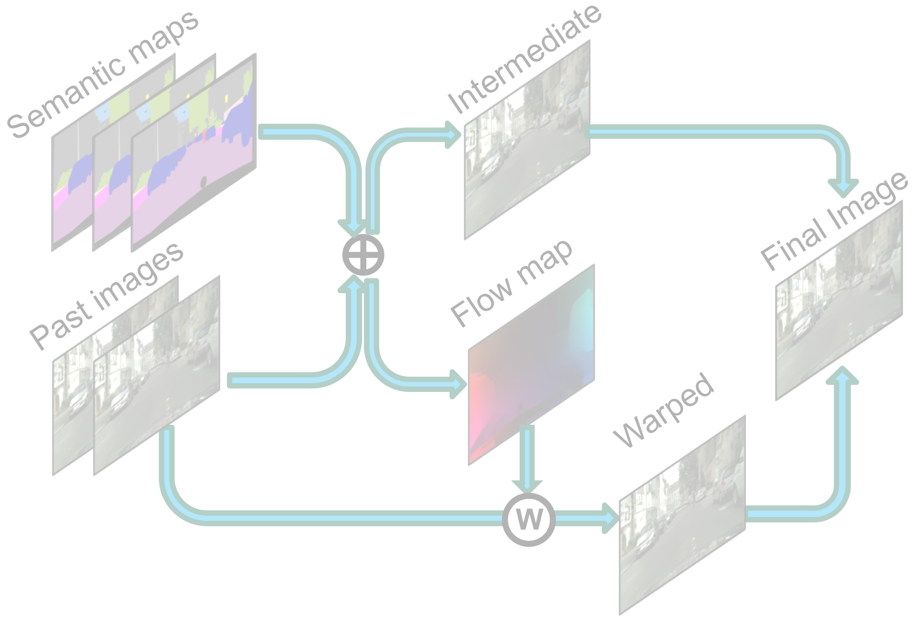
Sequential Generator



@ctnzr

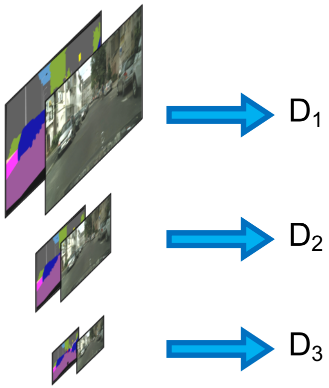
OUR METHOD

Sequential Generator

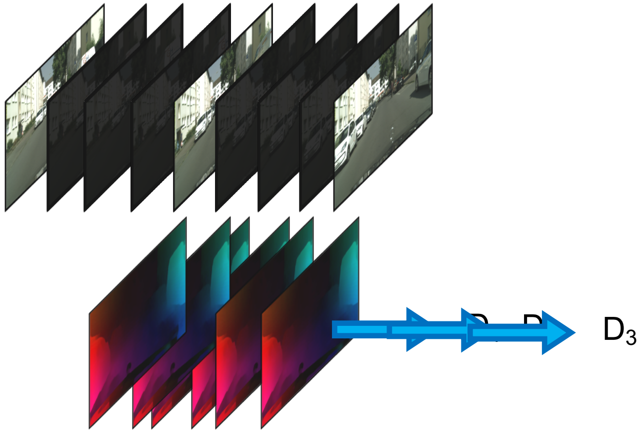


Multi-scale Discriminators

Image Discriminator



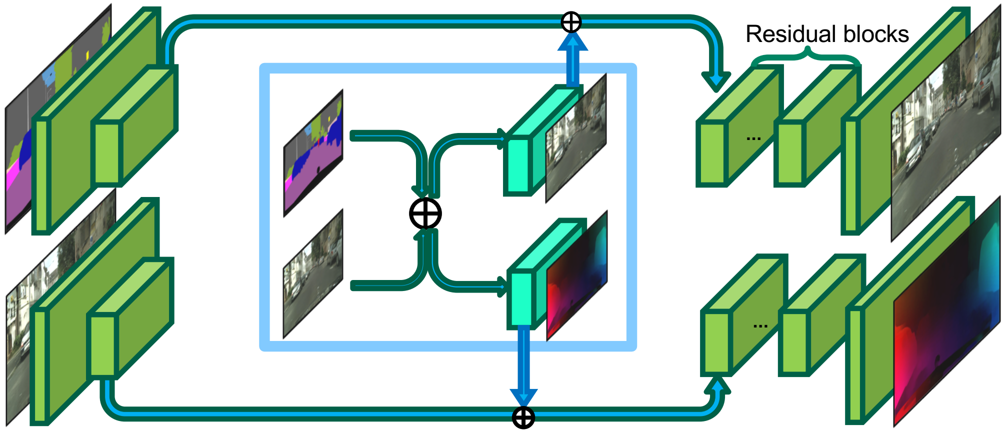
Video Discriminator



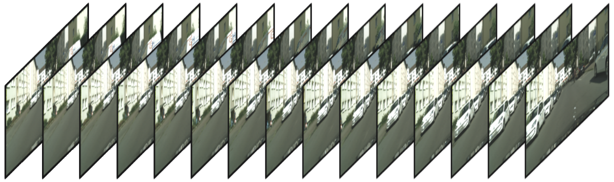
OUR METHOD

Spatio-temporally Progressive Training

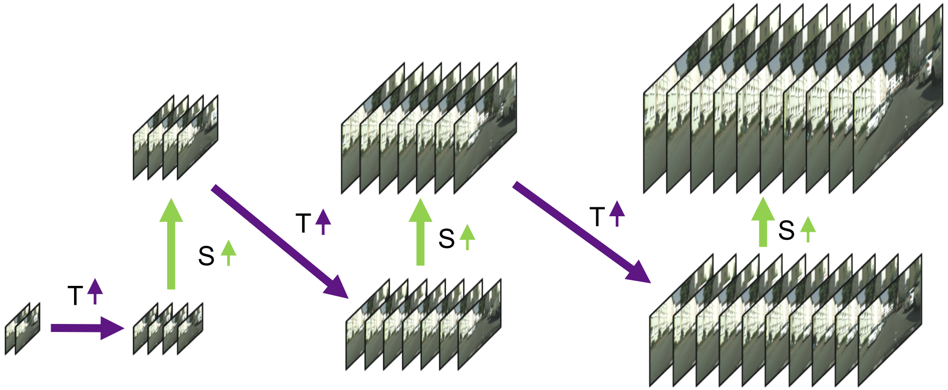
Spatially progressive



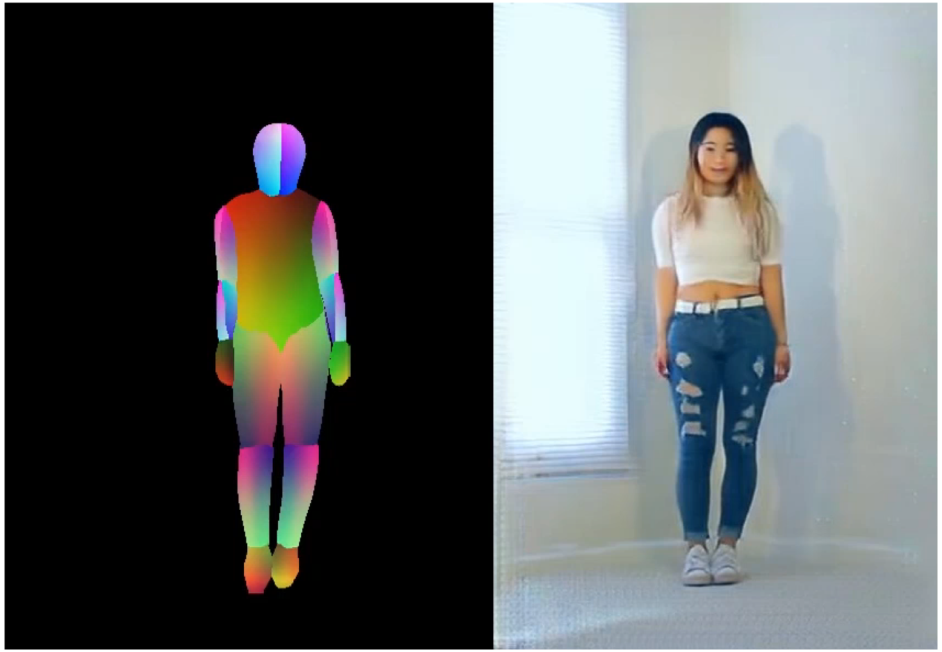
Temporally progressive

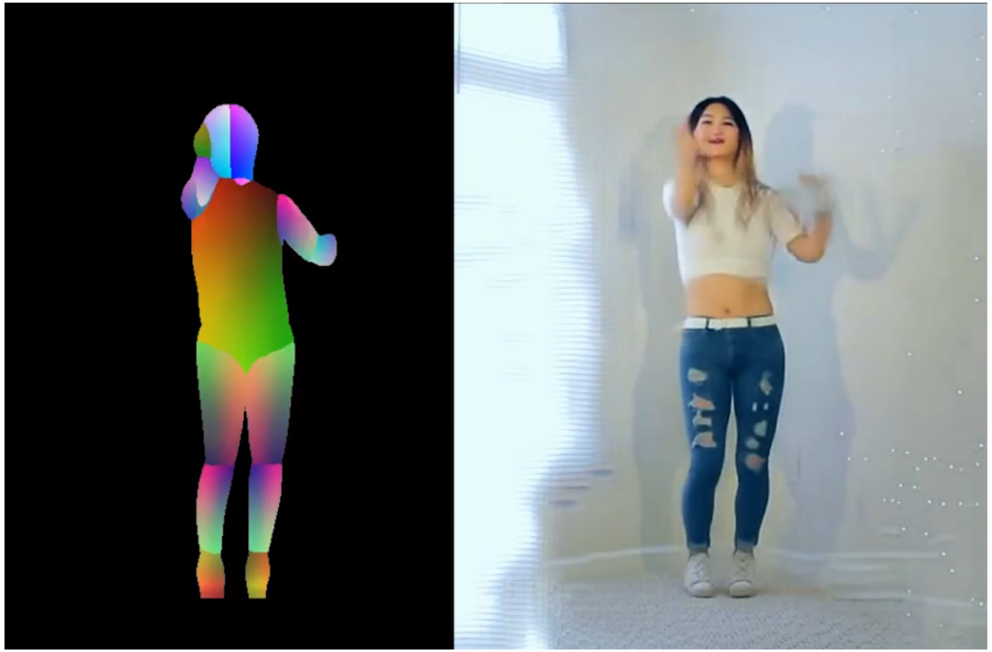


Alternating training



RESULTS: POSE-TO-BODY





AND IT RUNS IN REAL-TIME

https://bit.ly/vid2vid

SEMI-SUPERVISED LEARNING FOR NLP

https://github.com/NVIDIA/sentiment-discovery

Converge language model on 40 GB of text in 4 hours

Original 1 GPU, FP32 run took 1 month

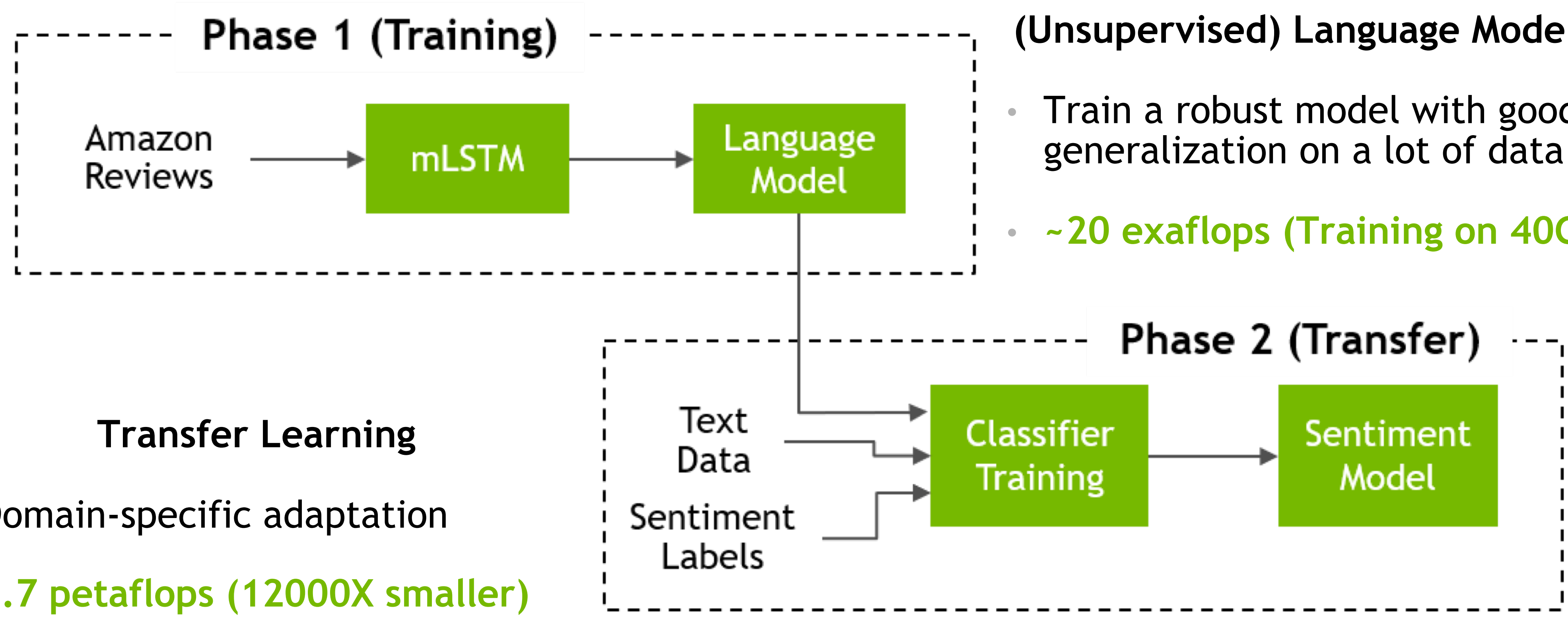
Using mixed precision arithmetic on 128 V100 GPUs

Transfer language model to sentiment task

Puri et al., https://arxiv.org/abs/1808.01371

Kant et al., https://arxiv.org/abs/1812.01207

LANGUAGE MODEL PRETRAINING & TRANSFER



(Unsupervised) Language Modeling

- Train a robust model with good generalization on a lot of data
- ~20 exaflops (Training on 40GB)

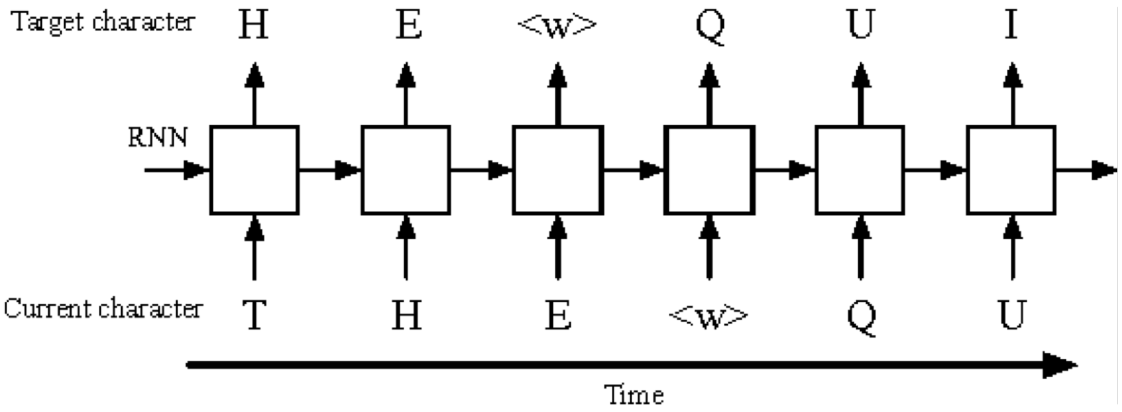
Transfer Learning

- Domain-specific adaptation
- 1.7 petaflops (12000X smaller)

UNSUPERVISED TRAINING

- Train a large neural network via next character prediction
  - Model learns dynamics of language
  - NO LABELS NEEDED - Label is next character

- 40GB of sentiment-filled Amazon Review data



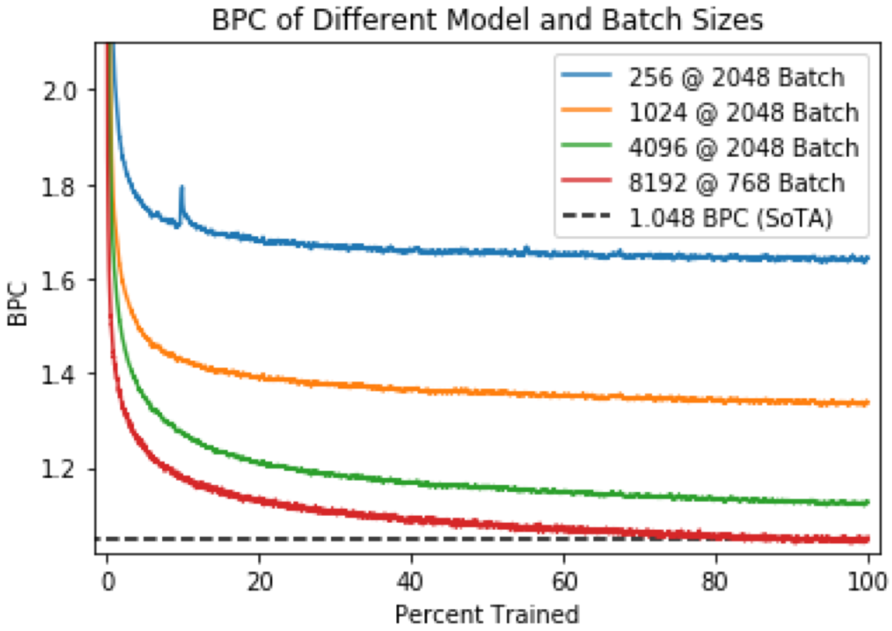
source

Sample Reviews
Shadows was an amazing book that caught my imagination instantly! It had love, brutality, adventure, and suspense that captivates your mind throughout the whole book.
the hooks were not chipped shipping was really fast nothing was broken all hooks were in package as described with all the sizes A+++++ thank you
Love this feeder. Heavy duty & capacity. Best feature is the large varmint guard. Definitely use a small lock or securing device on the battery housing latch. I gave 4 stars because several bolts were missing. Check contents b4 beginning.
The mp3 comes in Chinese!!! I DON'T KNOW THAT LANGUAGE, I AM ORDERING FOM USA. I DON'T UNDERSTAND ANYTHING AND I AM NOT ABLE TO CHANGE IT!!

@ctnzr

UNSUPERVISED TRAINING – LARGE MODELS

- Pretraining + transfer works with different model sizes
- Bigger better language model = better transfer
- Pretraining large models is expensive
- Scaling training is necessary for practicality

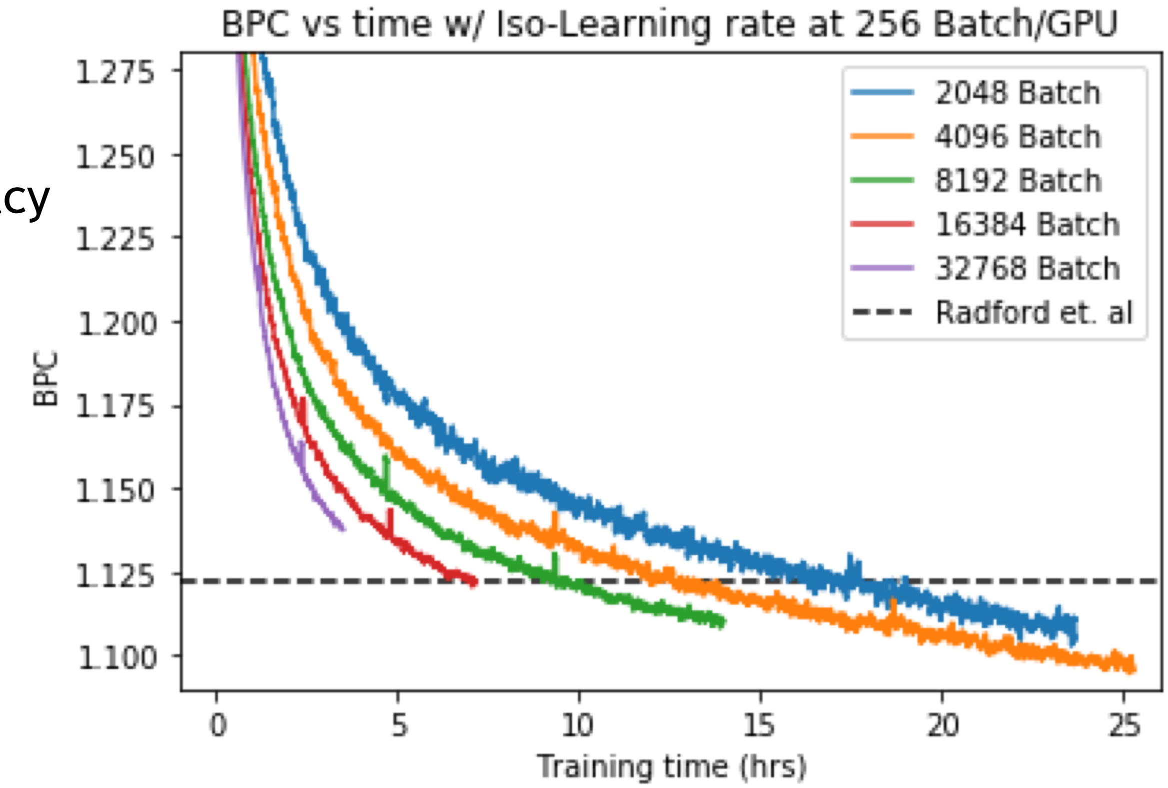


Hidden Size	FLOPS		BPC	SST	IMDB
	LM	Transfer			
256	1.14e17	3.19e12	1.541	53.2	62.2
1024	1.35e18	1.14e14	1.263	81.8	76.2
4096	**2.01e19**	**1.67e15**	**1.073**	**91.5**	**92.8**
8192	7.91e19	6.62e15	1.036	93.8	94.8

LARGE BATCH TRAINING

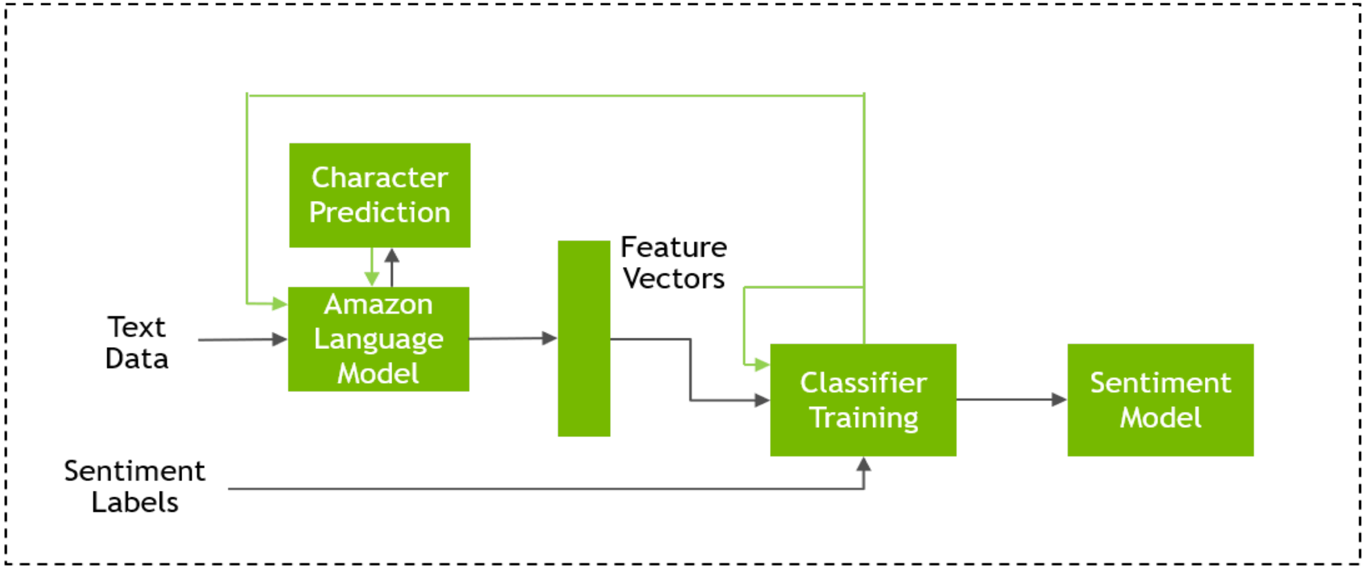
- Train with 32k batch size on 128 GPUs
- Converges with reasonable transfer accuracy in 3.5 hours

Batch	GPU	Iters	Ep	hrs	BPC	SST	IMDB
2048	8	100k	1.4	23.7	1.102	90.6	92.1
4096	16	100k	2.7	25.3	1.090	90.6	92.7
8192	32	55k	3.0	14.0	1.104	91.2	92.3
16384	64	28k	3.0	7.1	1.116	90.3	92.3
32768	128	14k	3.0	3.5	1.132	90.1	90.4



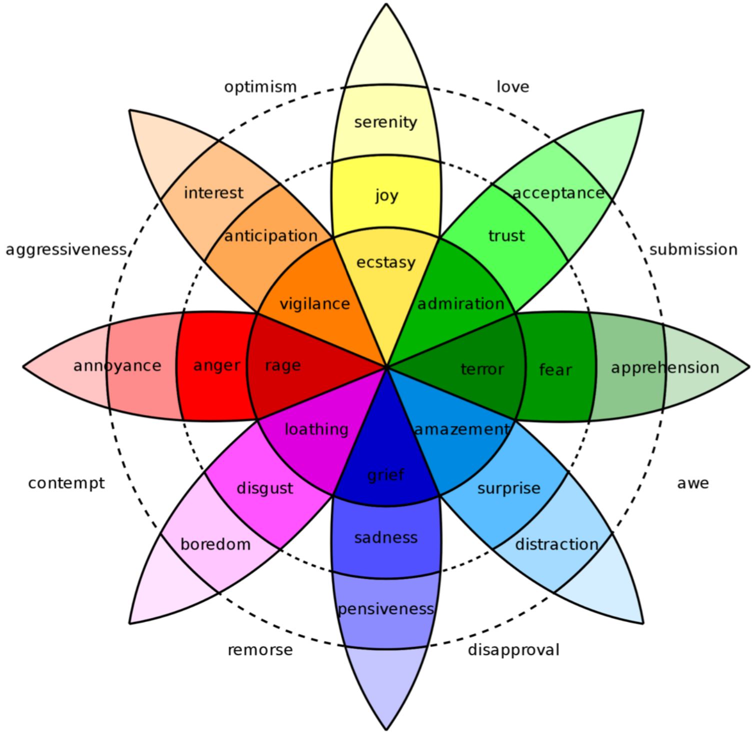
TRANSFER LEARNING & FINE TUNING

1. Initialize model with weights from pretraining
2. Model is used to featurize bodies of text
3. Binary Sentiment Classifier is trained on text features, while adjusting language model
4. Output Model: language model base + classifier on top

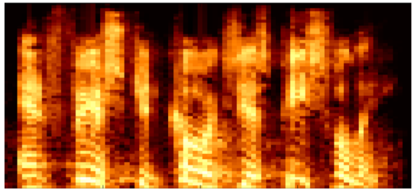
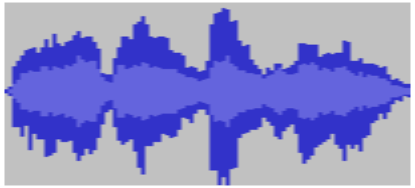


FINETUNING RESULTS ON SEMEVAL

- Finetuning and transfer works well with the transformer achieving even better results across numerous tasks than the mLSTM
- State of the art results for Plutchik emotion classification on the SemEval challenge
- Custom models for specific purposes,
  like NVIDIA social marketing



http://nv-adlr.github.io/WaveGlow

Hello, world! → TacoTron →  mel-spectrogram → WaveGlow →  audio samples

A new vocoder for speech synthesis built on a flow based generative model

Fast, completely parallel inference procedure

150X real-time on one V100 GPU

FLOW BASED GENERATIVE MODELS

Laurent Dinh, et al., 2014, 2016

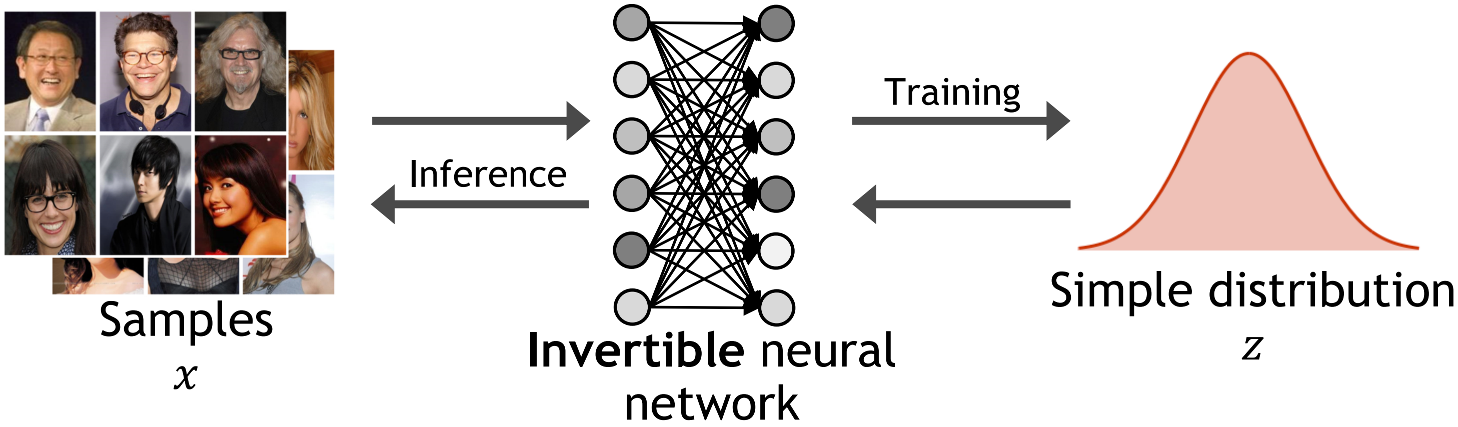
Generative model:

sample x from an unknown distribution: $x \sim p^*(x)$

Easy, if we only knew the distribution!

x is a multidimensional vector

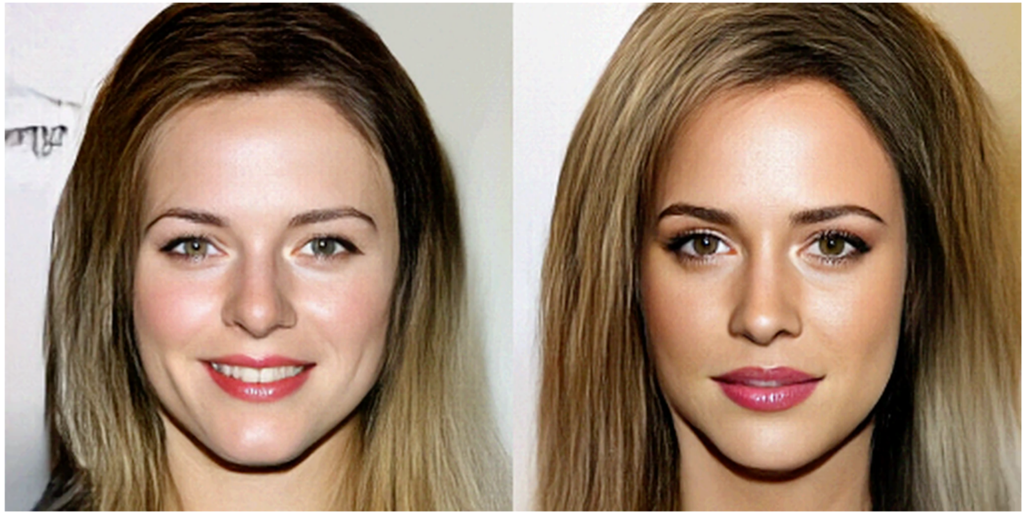
Flow based model:



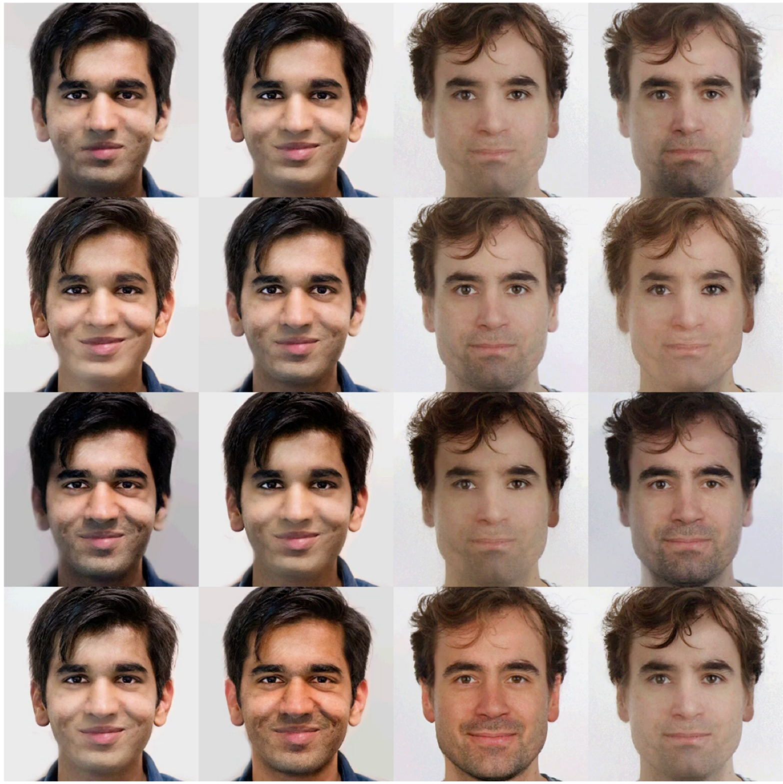
This is a change of variables!

OPENAI GLOW MODEL

https://blog.openai.com/glow/



Random samples
from GLOW model
trained on celebrities



Interpolating in latent space

INVERTIBLE NEURAL NETWORK??

By construction…

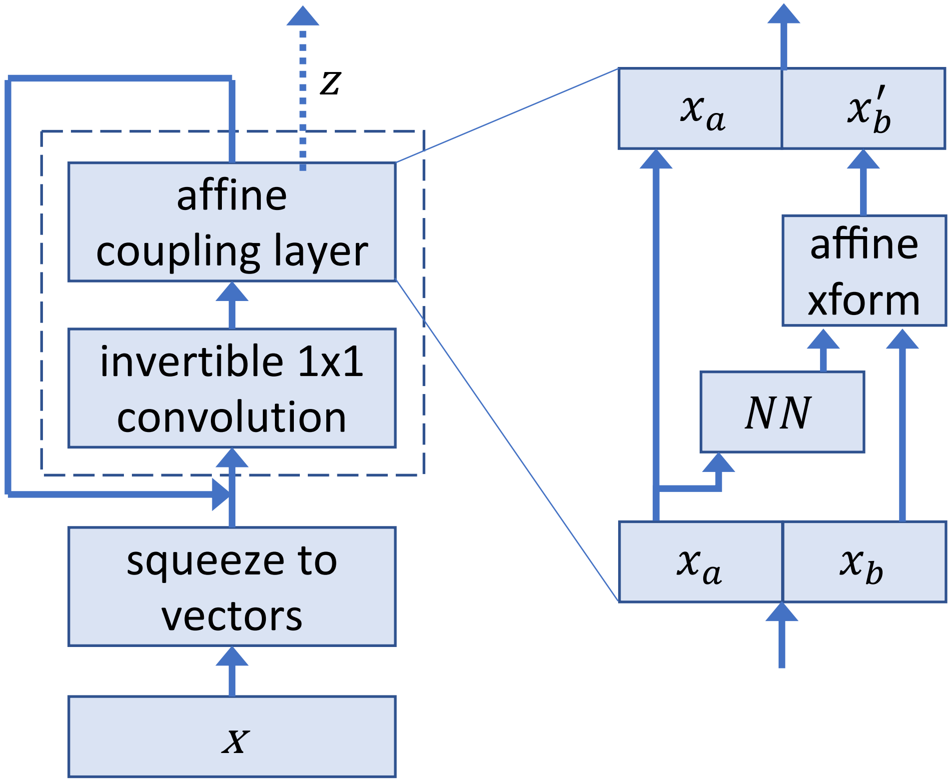
GLOW network built from two stages

Affine coupling layer

- Splits input channels in two
- Applies arbitrary network to half
- Computes affine xform for other half

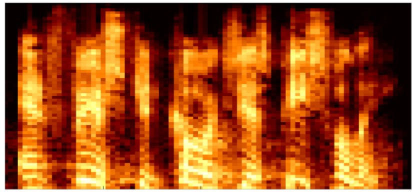
Invertible 1x1 convolutions

- Rotations mix information between channels



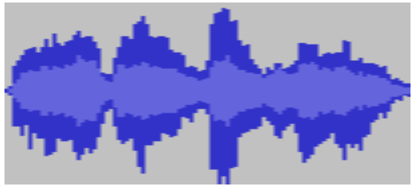
FROM GLOW TO WAVEGLOW

https://nv-adlr.github.io/WaveGlow



mel-spectrogram

vocoder



audio samples

Speech synthesis as sampling:

Sample from distribution of audio samples, conditioned on mel-spectrogram

Best speech synthesis today is autoregressive (sequential inference is hard at 22 kHz!)

Or has unstable training procedures (like student/teacher)

GLOW models are not autoregressive, and have a simple, stable training process

WaveGlow inverts mel-spectrograms at 2500 kHz on 1 GPU

CONCLUSION

This is a Golden Age for deep learning applications

Semi-supervised learning gives us new tools for DL applications

Text, Audio, Graphics

Using semi-supervised learning often requires us to change the way we train our models and collect data

But the rewards are great

Questions:
@ctnzr

NVIDIA®