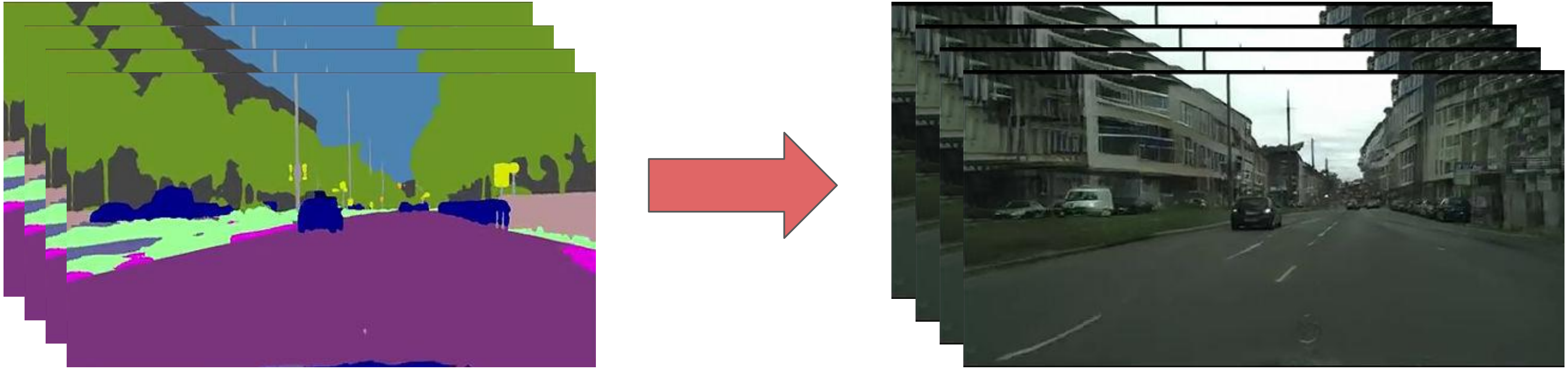
Article: Video-to-Video Synthesis

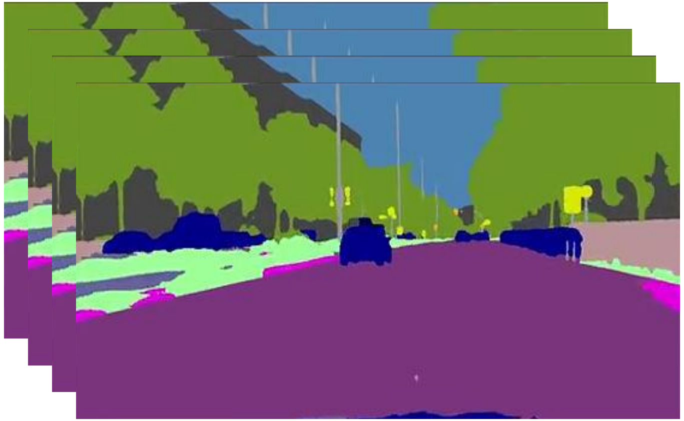
Speaker: Valeria Bubnova
27/09/2018

Goal

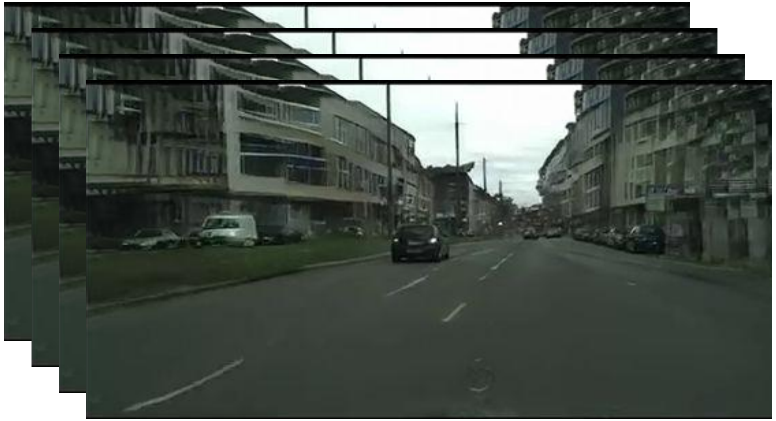


Goal

$$s_1^T = (s_1, ...s_T)$$



$$x_1^T = (x_1, ...x_T)$$



Goal

$$f(s_1^T) = \tilde{x}_1^T$$

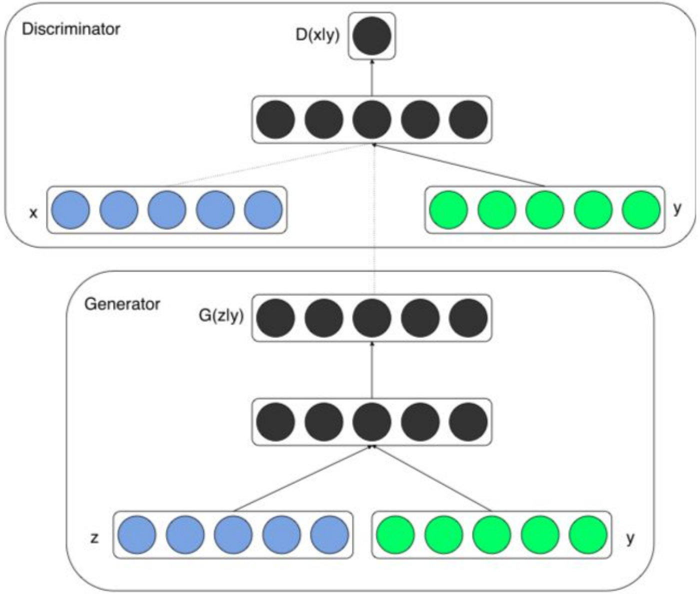
$$p(\tilde{x}_1^T | s_1^T) = p(x_1^T | s_1^T)$$

Goal

$$f(s_1^T) = \tilde{x}_1^T$$

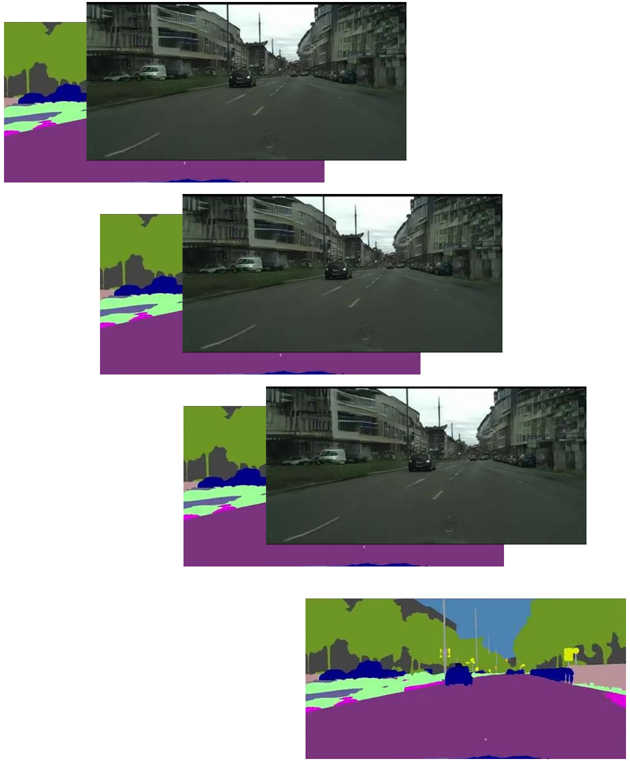
$$p(\tilde{x}_1^T | s_1^T) = p(x_1^T | s_1^T)$$

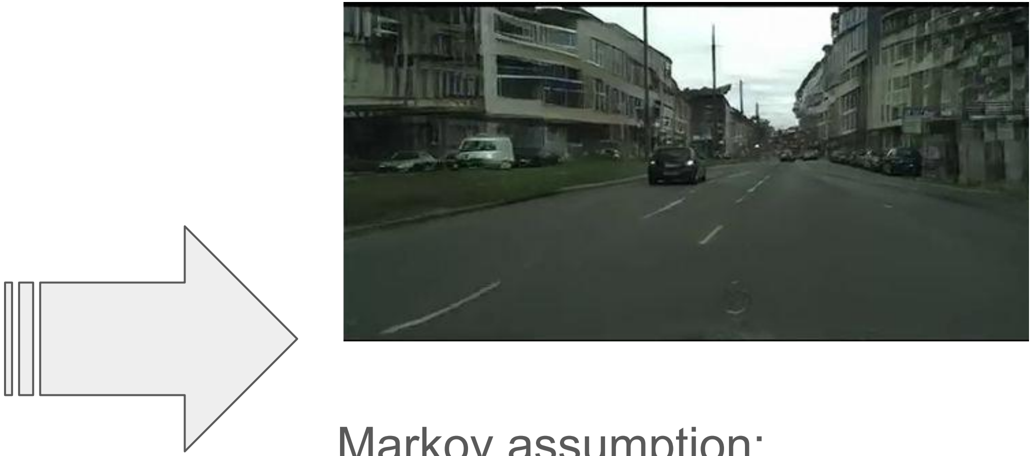
Conditional GAN



What is a video?

What is a video?





Markov assumption:

$$\tilde{x}_t = F(\tilde{x}_{t-L}^{t-1}, s_{t-L}^{t-1})$$

$$p(\tilde{x}_1^T | s_1^T) = \prod_{t=1}^{T} \boxed{p(\tilde{x}_t | \tilde{x}_{t-L}^{t-1}, s_{t-L}^{t})}$$

$$L = 2$$

What is an image?

What is an image?

Background
- Mostly saves its shape and details
- Moves on the canvas

Foreground
- May be moves a bit
- Changes
- Depends on the sourse

What is an image?

Background

- Mostly saves its shape and details
- Moves on the canvas

$$\tilde{h}_{B,t} = H_B(\tilde{x}_{t-L}^{t-1}, s_{t-L}^{t})$$

Foreground

- May be moves a bit
- Changes
- Depends on the sourse

$$\tilde{h}_{F,t} = H_F(s_{t-L}^{t})$$

How do images change?

Background

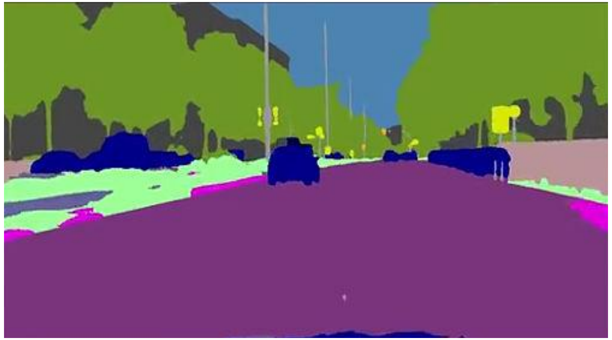
- Mostly saves its shape and details
- Moves on the canvas

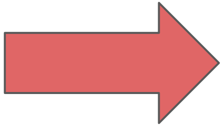
$$\tilde{h}_{B,t} = H_B(\tilde{x}_{t-L}^{t-1}, s_{t-L}^{t})$$

1) Optical flow

How do images change?

2) From the source





How do we get the final t-th image?

$$F(\tilde{x}_{t-L}^{t-1}, s_{t-L}^{t-1}) = (1 - \tilde{m}_t) \odot \tilde{w}_{t-1}(\tilde{x}_{t-1}) +$$

$$+ \tilde{m}_t \odot ((1 - m_{B,t}) \odot \tilde{h}_{F,t} + m_{B,t} \cdot \tilde{h}_{B,t})$$

How do we get the final t-th image?

Mask $\tilde{m}_t \in (0,1)^{n \times m}$

Image transformed by optical flow

$$F(\tilde{x}_{t-L}^{t-1}, s_{t-L}^{t-1}) = (1 - \tilde{m}_t) \odot \tilde{w}_{t-1}(\tilde{x}_{t-1}) +$$

$$+ \tilde{m}_t \odot ((1 - m_{B,t}) \odot \tilde{h}_{F,t} + m_{B,t} \cdot \tilde{h}_{B,t})$$

Ground truth background mask based on s

Newly generated images

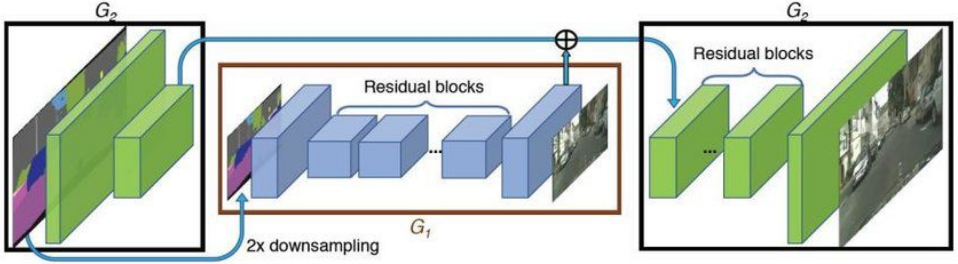
How do we get the final t-th image?

$$\tilde{w}_{t-1} = W(\tilde{x}_{t-L}^{t-1}, s_{t-L}^{t})$$

$$\tilde{m}_{t-1} = M(\tilde{x}_{t-L}^{t-1}, s_{t-L}^{t})$$

$$\tilde{h}_{t-1} = H(\tilde{x}_{t-L}^{t-1}, s_{t-L}^{t})$$

Same NN, differ in last layer only:
Residual Network Architecture by Wang et al.



#Pix2Pix

Which loss?

Jensen-Shannon divergence

$$\max_{D} \min_{G} E_{(\mathbf{x}_1^T, \mathbf{s}_1^T)}[\log D(\mathbf{x}_1^T, \mathbf{s}_1^T)] + E_{\mathbf{s}_1^T}[\log(1 - D(G(\mathbf{s}_1^T), \mathbf{s}_1^T))]$$

Which loss?

$$D_i :$$
$$(x_t, s_t) \rightarrow 1$$
$$(\tilde{x}_t, s_t) \rightarrow 0$$

$$D_v :$$
$$(x_{t-K}^{t-1}, w_{t-K}^{t-2}) \rightarrow 1$$
$$(\tilde{x}_{t-K}^{t-1}, w_{t-K}^{t-2}) \rightarrow 0$$

Architecture: PatchGan

Which loss?

$$\boxed{\min_{F}\Big(\max_{D_I}\mathcal{L}_I(F, D_I) + \max_{D_V}\mathcal{L}_V(F, D_V)\Big) + \lambda_W\mathcal{L}_W(F)}$$

Which loss?

$$\boxed{\min_F \Big(\max_{D_I} \mathcal{L}_I(F, D_I) + \max_{D_V} \mathcal{L}_V(F, D_V)\Big) + \lambda_W \mathcal{L}_W(F)}$$

$$\mathcal{L}_I = E_{\phi_I(\mathbf{x}_1^T, \mathbf{s}_1^T)}[\log D_I(\mathbf{x}_i, \mathbf{s}_i)] + E_{\phi_I(\tilde{\mathbf{x}}_1^T, \mathbf{s}_1^T)}[\log(1 - D_I(\tilde{\mathbf{x}}_i, \mathbf{s}_i))]$$

$$\begin{aligned}\mathcal{L}_V = E_{\phi_V(\mathbf{w}_1^{T-1}, \mathbf{x}_1^T, \mathbf{s}_1^T)}[\log D_V(\mathbf{x}_{i-K}^{i-1}, \mathbf{w}_{i-K}^{i-2})] +& \\ + E_{\phi_V(\mathbf{w}_1^{T-1}, \tilde{\mathbf{x}}_1^T, \mathbf{s}_1^T)}[\log(1 - D_V(\tilde{\mathbf{x}}_{i-K}^{i-1}, \mathbf{w}_{i-K}^{i-2}))]&\end{aligned}$$

$$\mathcal{L}_W = \frac{1}{T-1} \sum_{t=1}^{T-1} \Big(\|\tilde{\mathbf{w}}_t - \mathbf{w}_t\|_1 + \|\tilde{\mathbf{w}}_t(\mathbf{x}_t) - \mathbf{x}_{t+1}\|_1 \Big)$$

Details

- ADAM, 40 epochs
- NVIDIA 8 V100 16GB (DGX1 machine)
- Increasing resolution (up to 2048*1024)
- 10 days (for 2K)

Multiple Outputs for Edge-to-Face

Datasets

- Cityscape
- Apploscape
- Face video dataset
- Danse Video Dataset

Metrics

- Fréchet Inception Distance

$$\|\mu - \tilde{\mu}\|^2 + \mathbf{Tr}\left(\Sigma + \tilde{\Sigma} - 2\sqrt{\Sigma\tilde{\Sigma}}\right)$$

- Human Preference Score

Results:

Table 1: Comparison between competing video-to-video synthesis approaches on Cityscapes.

Fréchet Inception Distance	I3D	ResNeXt
pix2pixHD	5.57	0.18
COVST	5.55	0.18
vid2vid (ours)	**4.66**	**0.15**

Human Preference Score	short seq.	long seq.
vid2vid (ours) / pix2pixHD	**0.87** / 0.13	**0.83** / 0.17
vid2vid (ours) / COVST	**0.84** / 0.16	**0.80** / 0.20

Table 2: Ablation study. We compare the proposed approach to its three variants.

Human Preference Score	
vid2vid (ours) / no background-foreground prior	**0.80** / 0.20
vid2vid (ours) / no conditional video discriminator	**0.84** / 0.16
vid2vid (ours) / no flow warping	**0.67** / 0.33

Table 3: Comparison between future video prediction methods on Cityscapes.

Fréchet Inception Distance	I3D	ResNeXt
PredNet	11.18	0.59
MCNet	10.00	0.43
vid2vid (ours)	**3.44**	**0.18**

Human Preference Score	
vid2vid (ours) / PredNet	**0.92** / 0.08
vid2vid (ours) / MCNet	**0.98** / 0.02

Result

Video:

https://tcwang0509.github.io/vid2vid/paper_gifs/cityscapes_comparison.gif

https://tcwang0509.github.io/vid2vid/paper_gifs/apollo.gif

https://tcwang0509.github.io/vid2vid/paper_gifs/face.gif

https://tcwang0509.github.io/vid2vid/paper_gifs/pose.gif

Source:

Video-to-Video Synthesis

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

(Submitted on 20 Aug 2018)

https://arxiv.org/abs/1808.06601