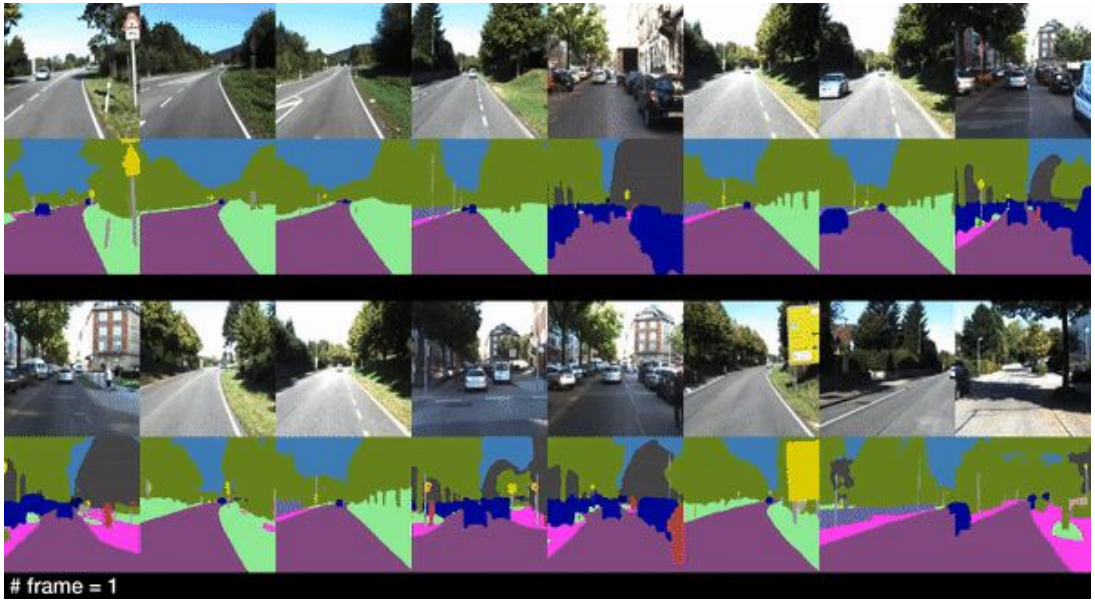
ICLR

Revisiting Hierarchical Approach For Persistent Long-term Video Prediction

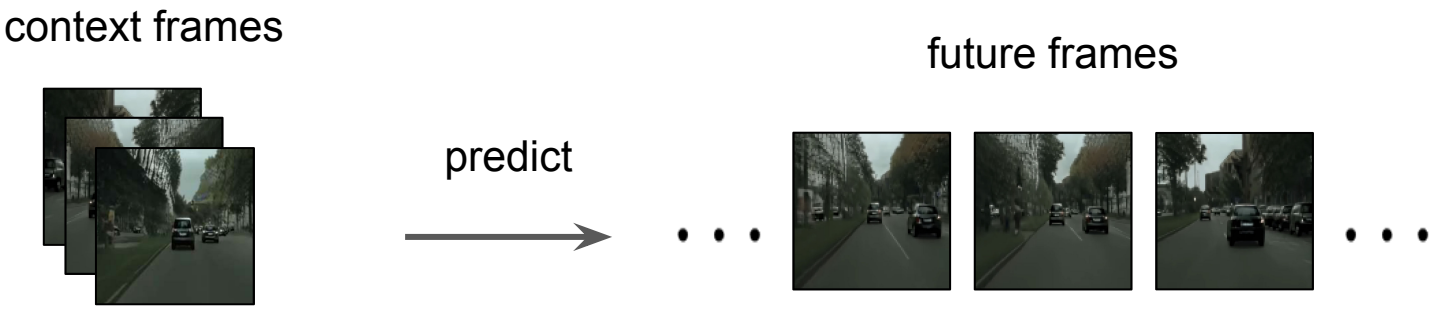
[1]Wonkwang Lee, [1]Whie Jung, [2]Han Zhang, [2]Ting Chen, [2]Jing Yu Koh,
[3]Thomas Huang, [4]Hyunsuk Yoon, [3,5]Honglak Lee, and [1]Seunghoon Hong



1 KAIST 2 Google Research 3 University of Michigan 4 Moloco 5 LG AI Research

Video Prediction

Predicting future frames given a few context frames as inputs.



$$\mathbf{x}_{1:C} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_C\}$$

$$\mathbf{x}_{C+1:T} = \{\mathbf{x}_{C+1}, \mathbf{x}_{C+2}, \cdots, \mathbf{x}_T\}$$

Challenges

Spatio-temporal variations and uncertainties in video sequences.

- Complex **structures** and **appearances** in the high-dimensional image data.
- Inherent **stochasticity** and **recurrency** in the real-world dynamics.

context frame

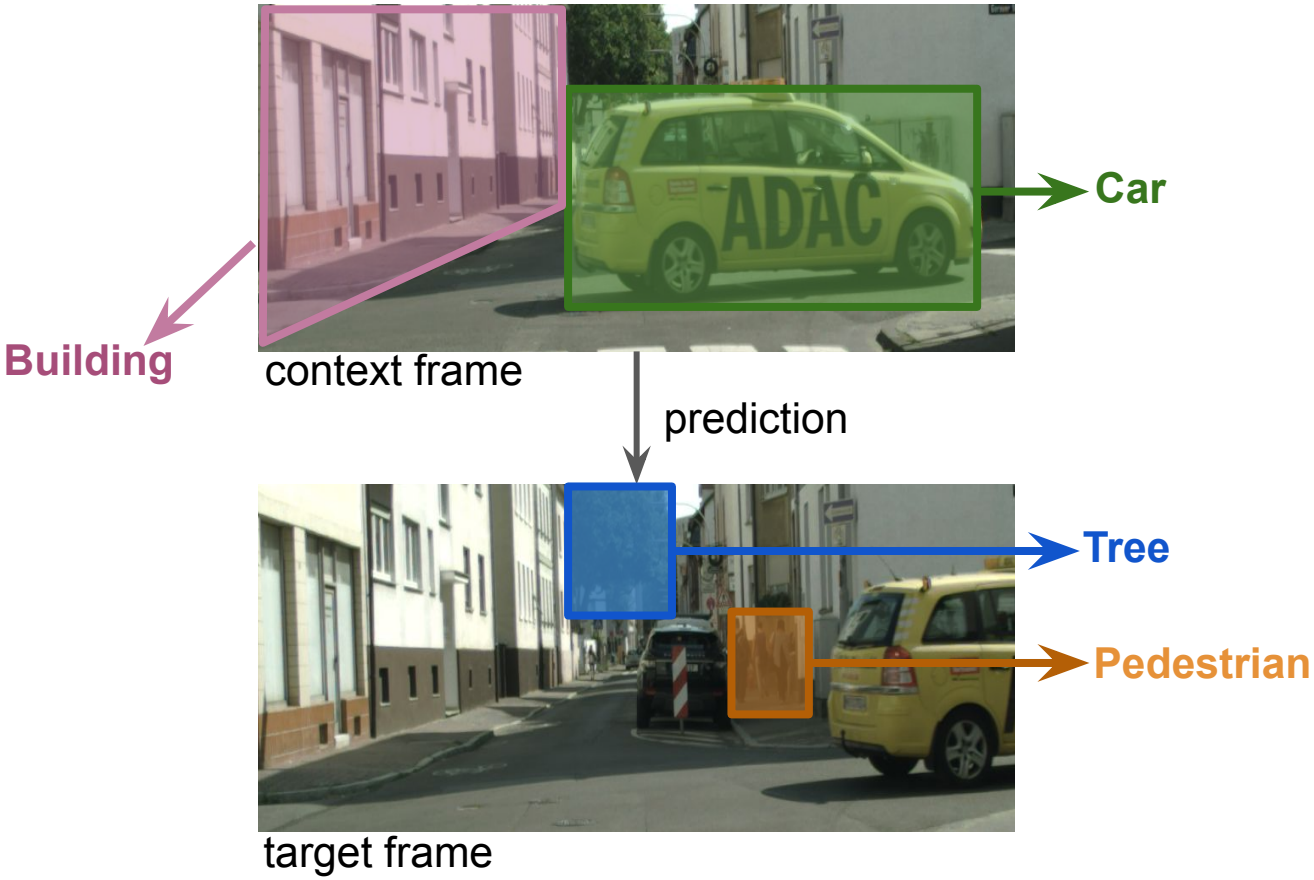
prediction

target frame

Challenges

Spatio-temporal variations and uncertainties in video sequences.

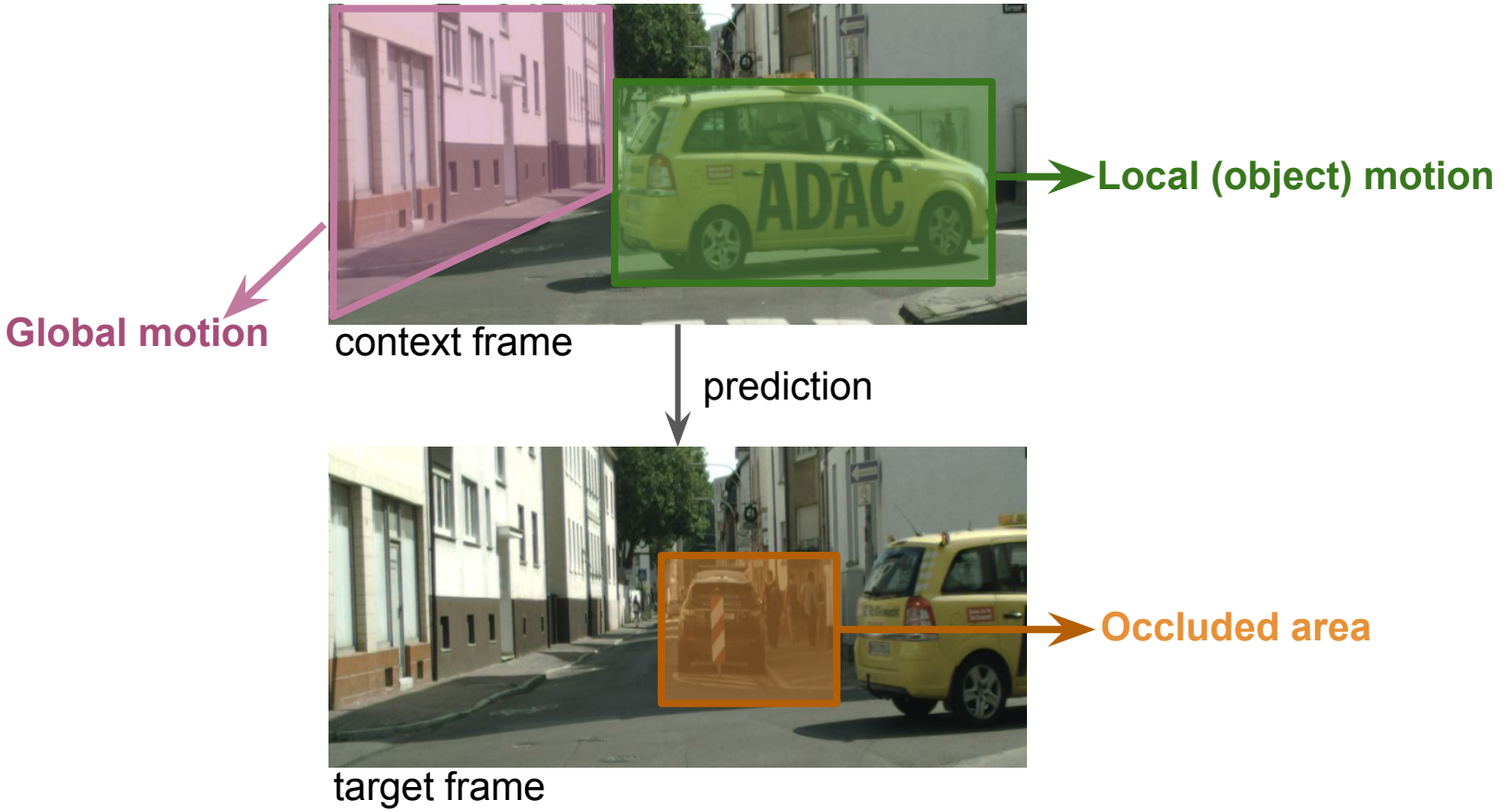
- Complex **structures** and **appearances** in the high-dimensional image data.
- Inherent **stochasticity** and **recurrency** in the real-world dynamics.



Challenges

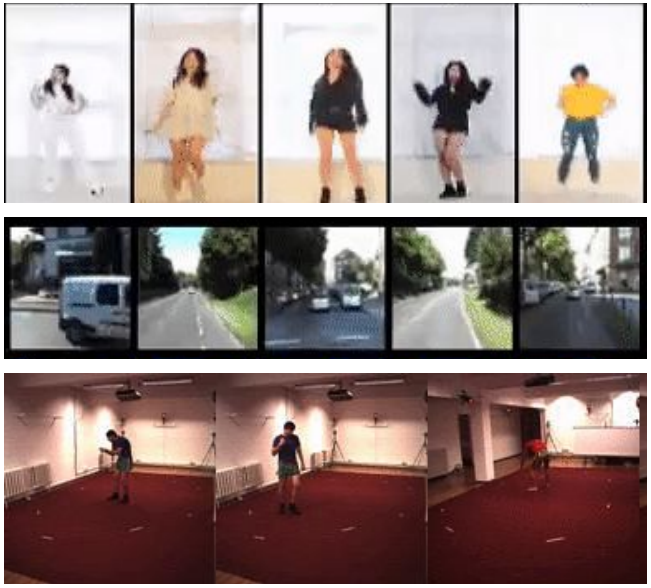
Spatio-temporal variations and uncertainties in video sequences.

- Complex **structures** and **appearances** in the high-dimensional image data.
- Inherent **stochasticity** and **recurrency** in the real-world dynamics.

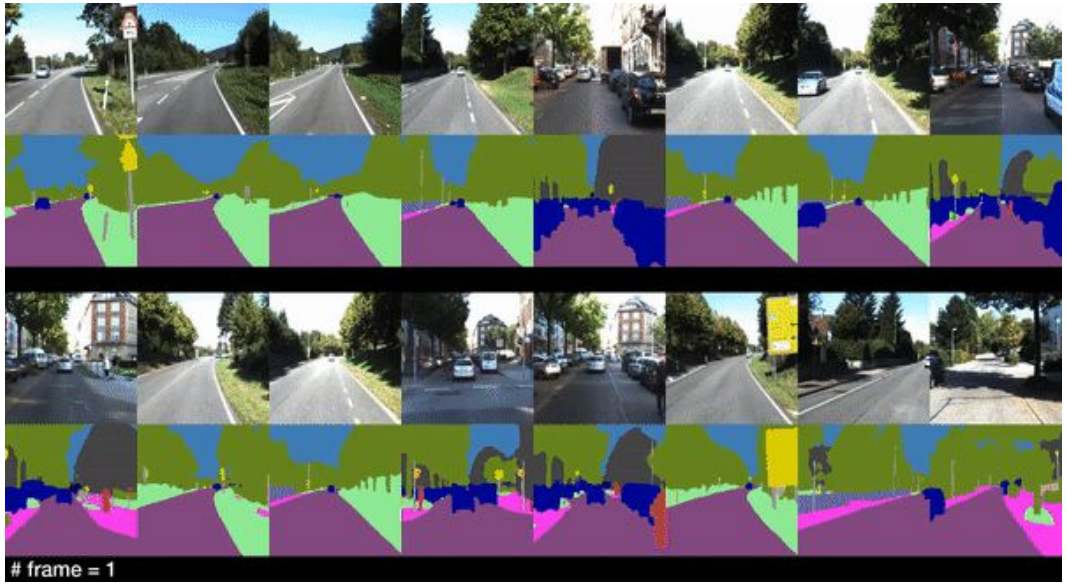


Prior Work Versus Ours

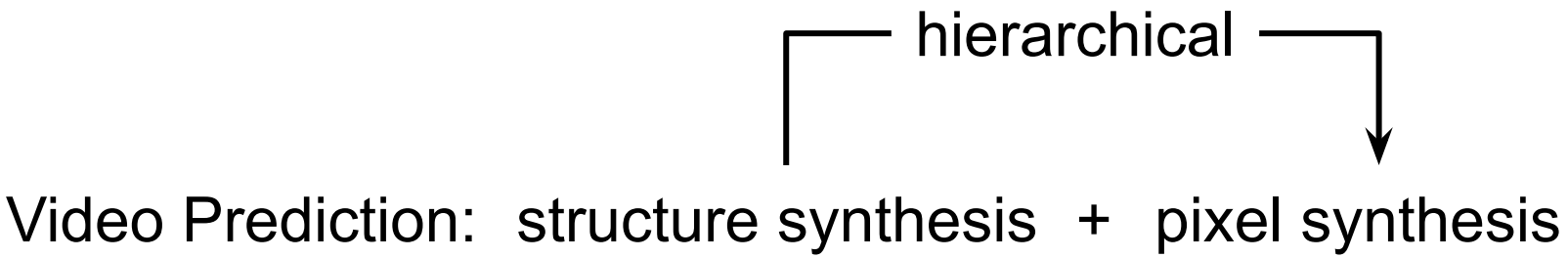
Prior Work



Ours



Method: Hierarchical Video Prediction Network



Method: Hierarchical Video Prediction Network

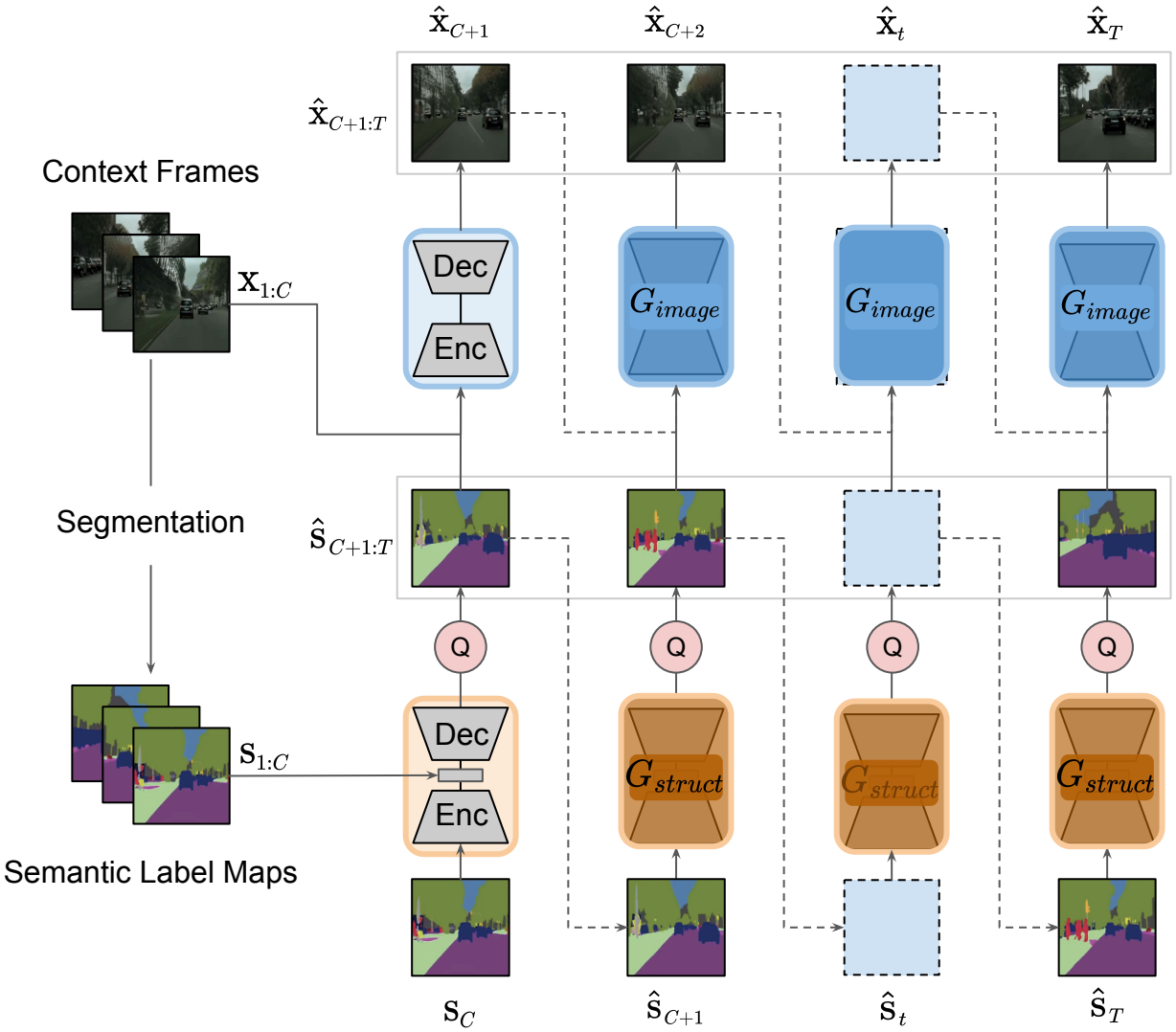


Image Generator

$p(x_{\leq T}|s_{\leq T}) = \prod_{t=1}^{T} p(x_t|x_{<t}, s_{\leq t})$

Hierarchical video prediction model:
- appearances of semantic structures
- dynamics of semantic structures

Structure Generator

$p(s_{\leq T}|z_{\leq T}) = \prod_{t=1}^{T} p(s_t|s_{<t}, z_{\leq t})$

Method: Hierarchical Video Prediction Network

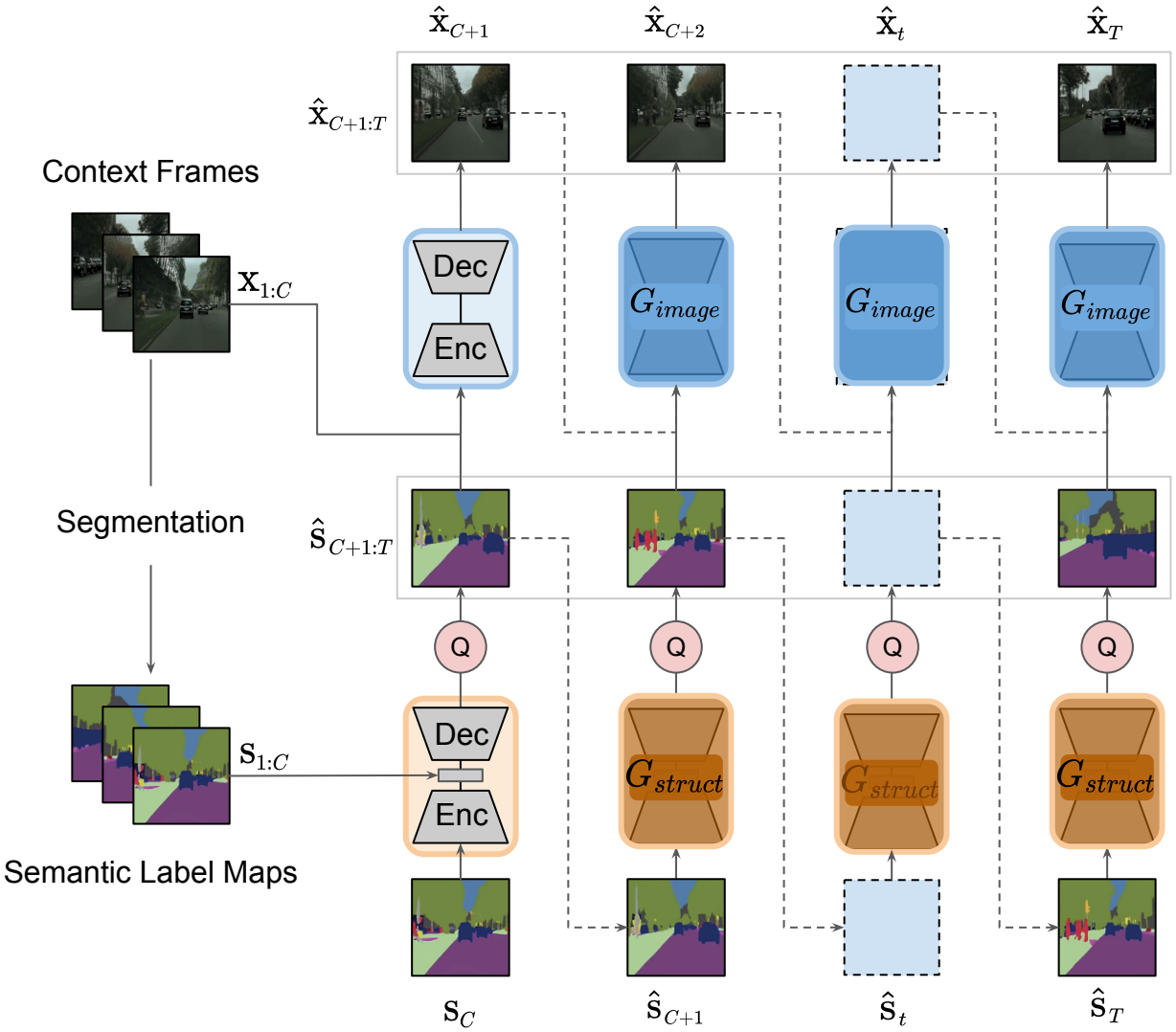


Image Generator

$p(x_{\leq T}|s_{\leq T}) = \prod_{t=1}^{T} p(x_t|x_{<t}, s_{\leq t})$

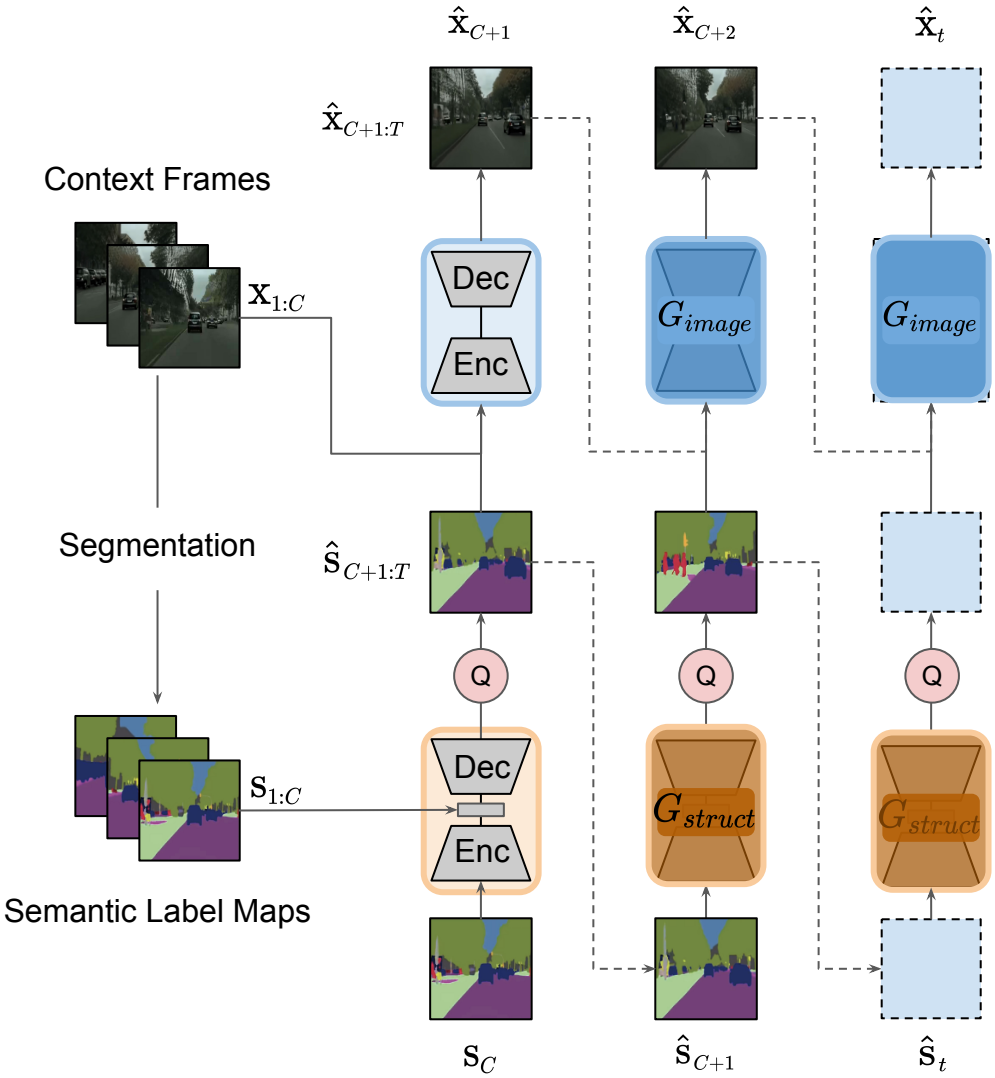
Video-to-video translation method:
- pixel-level temporal consistency
- pixel-level high-frequency details

Structure Generator

$p(s_{\leq T}|z_{\leq T}) = \prod_{t=1}^{T} p(s_t|s_{<t}, z_{\leq t})$

Stochastic recurrent neural network:
- structure-level stochasticity
- structure-level recurrency

Method: Robustness To Errors



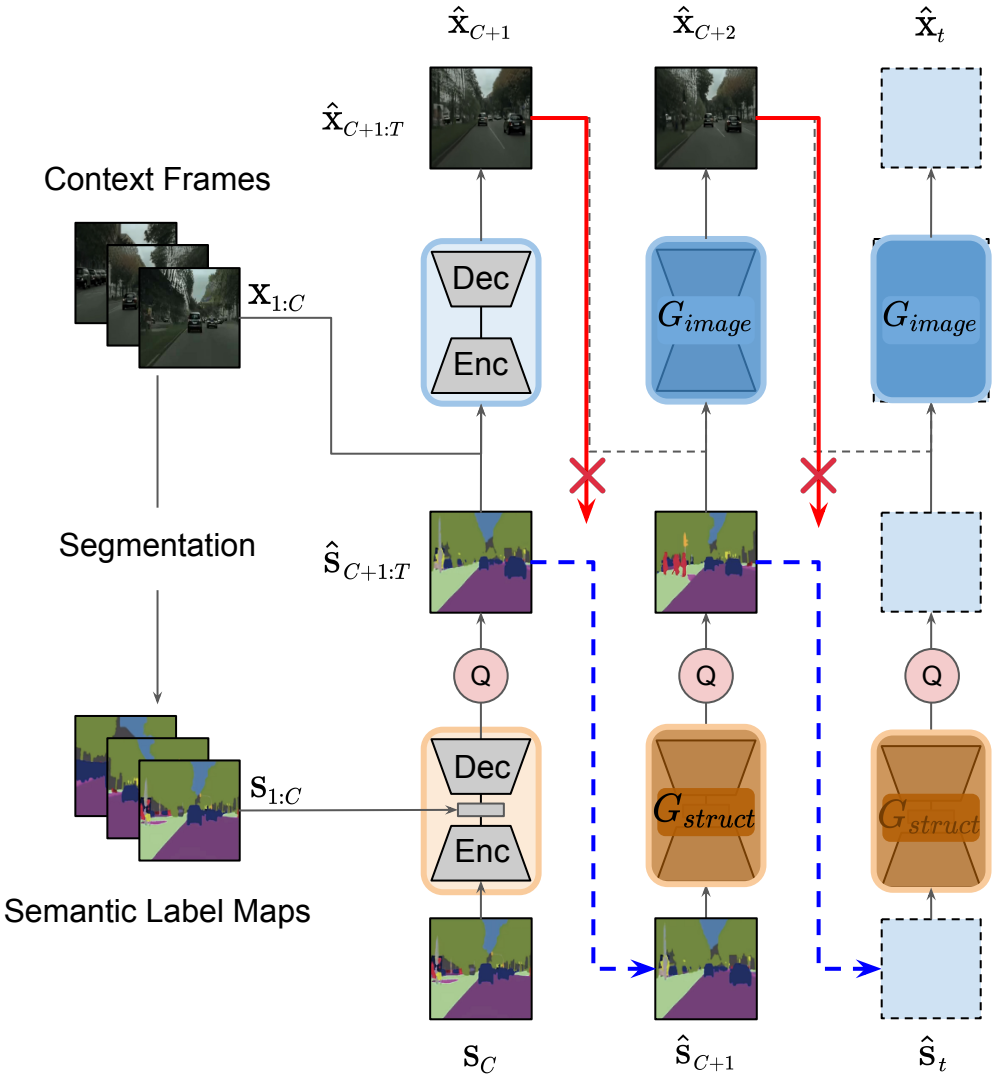
Independent unrolling loop for structures.

- Pixel-level errors do not propagate through the structure generator.

Discretization of structure representations.

- Structure-level blurriness is reduced.

Method: Robustness To Errors



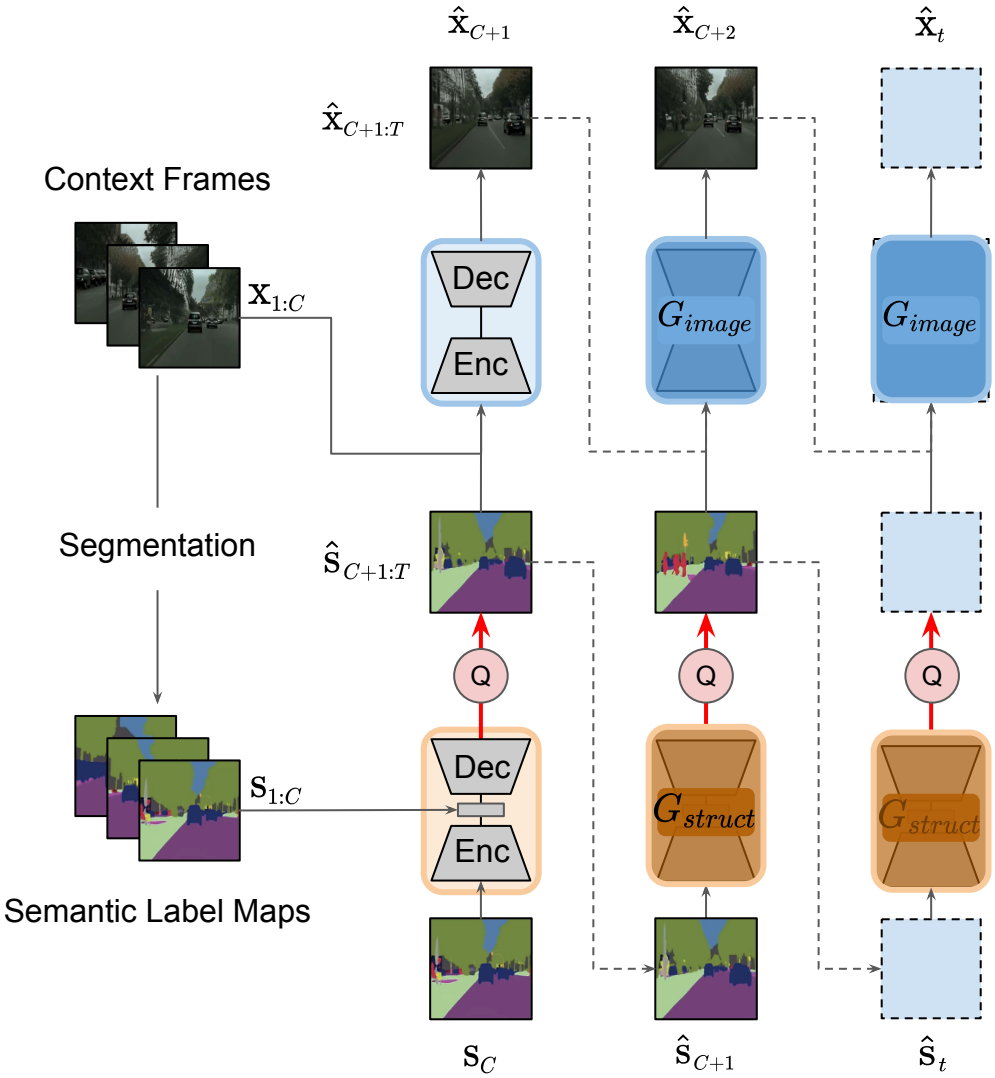
Independent unrolling loop for structures.

- Pixel-level errors do not propagate through the structure generator.

Discretization of structure representations.

- Structure-level blurriness is reduced.

Method: Robustness To Errors



Independent unrolling loop for structures.

- Pixel-level errors do not propagate through the structure generator.

Discretization of structure representations.

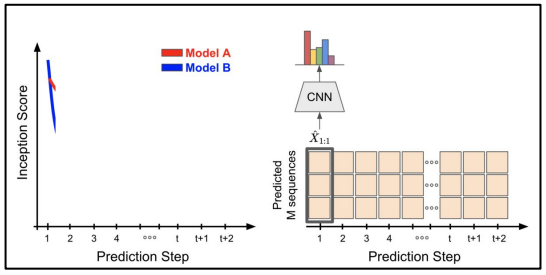
- Structure-level blurriness is reduced.

Baselines

Method	Hierarchical Estimation	Stochastic Estimation	Recurrent Estimation
SVG / SVG-extend (non-hierarchical)	x	o	o
Villegas et al. (deterministic)	o	x	o
Bayes-WD-SL (non-recurrent)	o	o	x
Ours	o	o	o

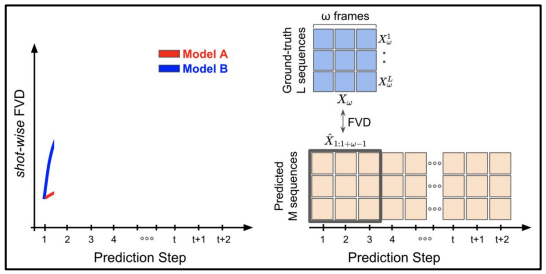
Metrics

- Inception Score *(higher-the-better)*



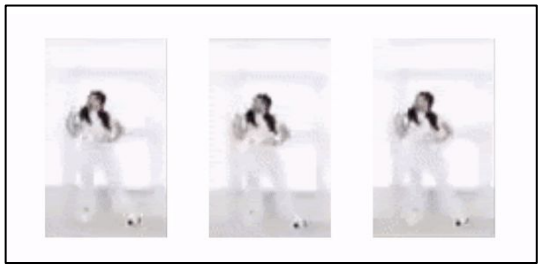
$$\text{Inception Score}(t) = \exp\left(\mathbb{E}_{x \sim \hat{X}_{t:t}}\left[D_{KL}(p(y|x)||p(y))\right]\right)$$

- *shot-wise* FVD *(lower-the-better)*



$$\text{shot-wise FVD}(t) = \text{FVD}(\hat{X}_{t:t+\omega-1}, X_\omega)$$

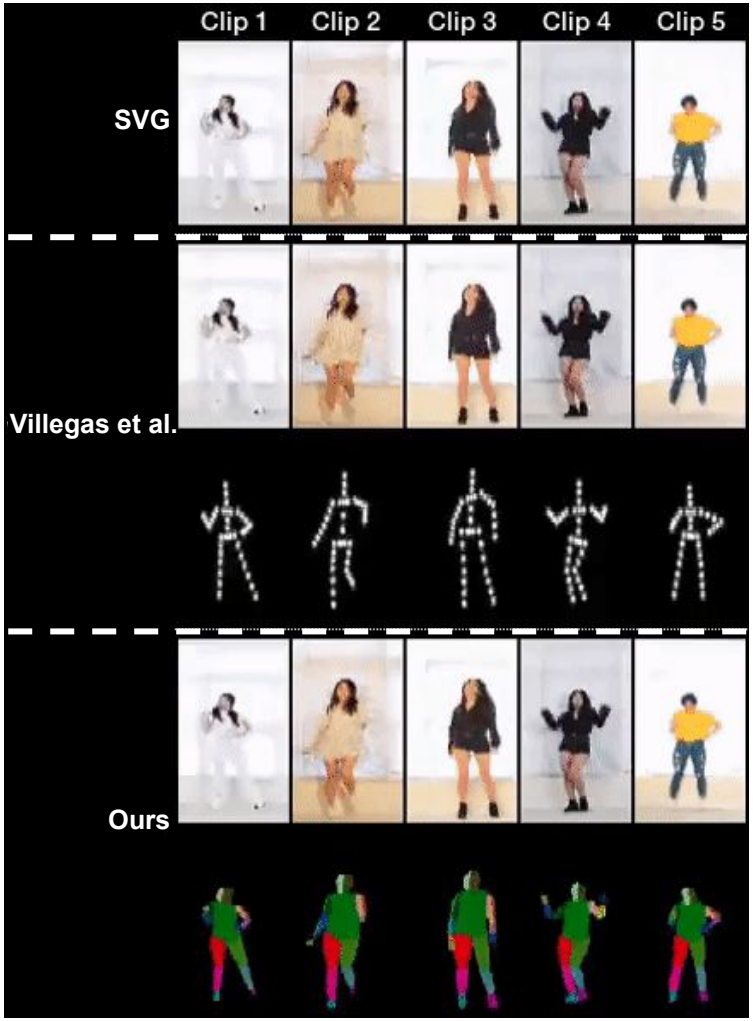
- Human Evaluation *(higher-the-better)*



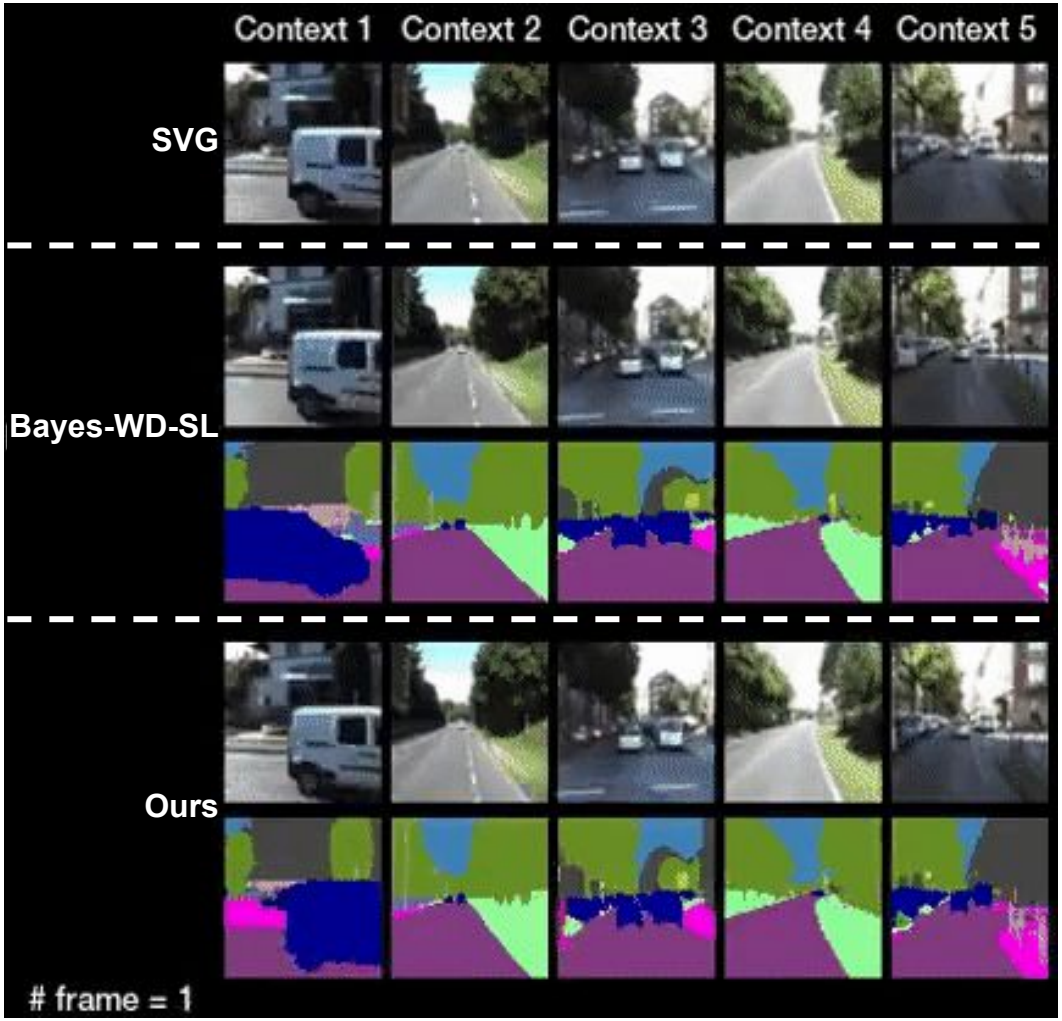
Q: which video do you prefer most?

Results: Qualitative (64x64)

Human Dancing

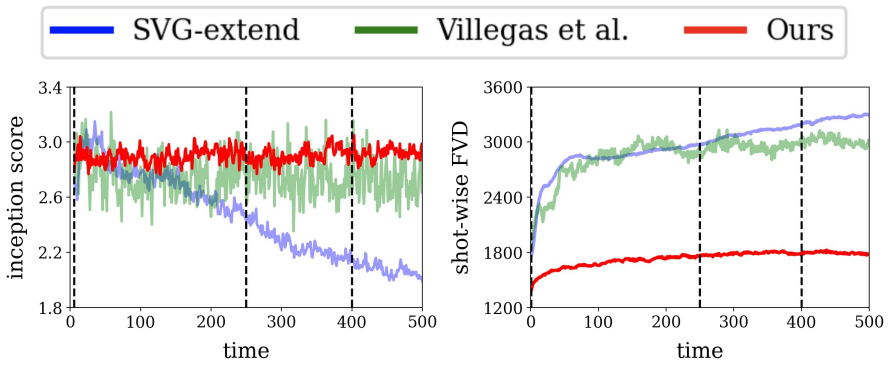


KITTI Benchmark



Results: Quantitative (64x64)

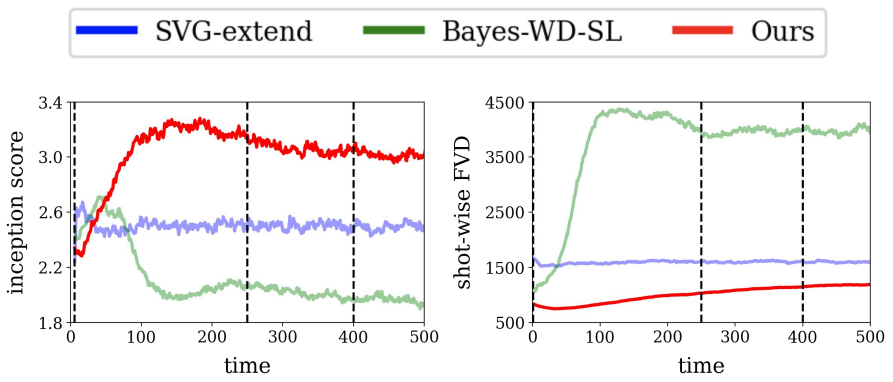
Human Dancing



Model	t=1	t=250	t=400
SVG-Extend	3.9	2.3	3.1
Villegas et al.	6.6	9.9	9.1
Ours	**89.5**	**87.8**	**87.8**

Human evaluation (most-preferred ratio)

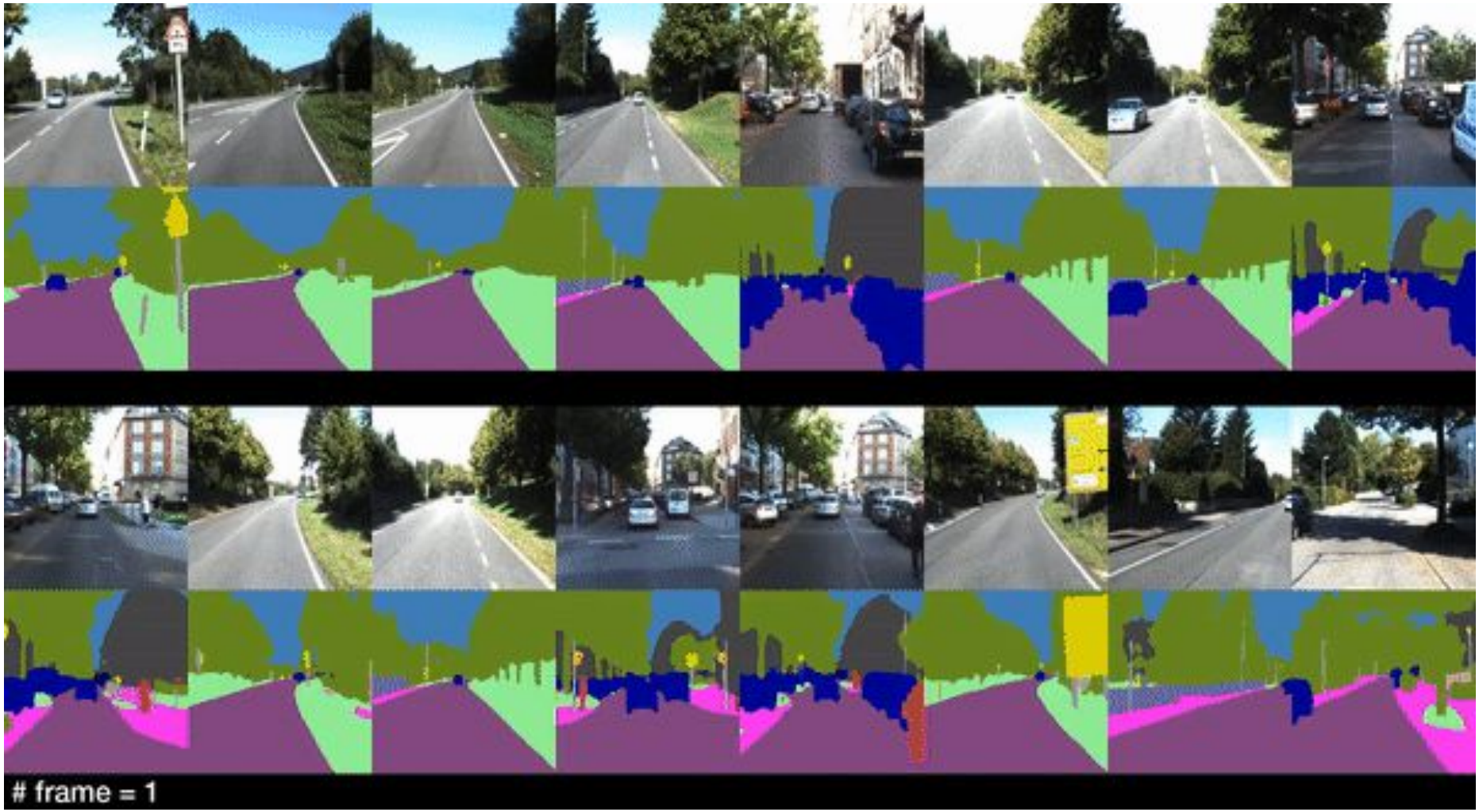
KITTI Benchmark



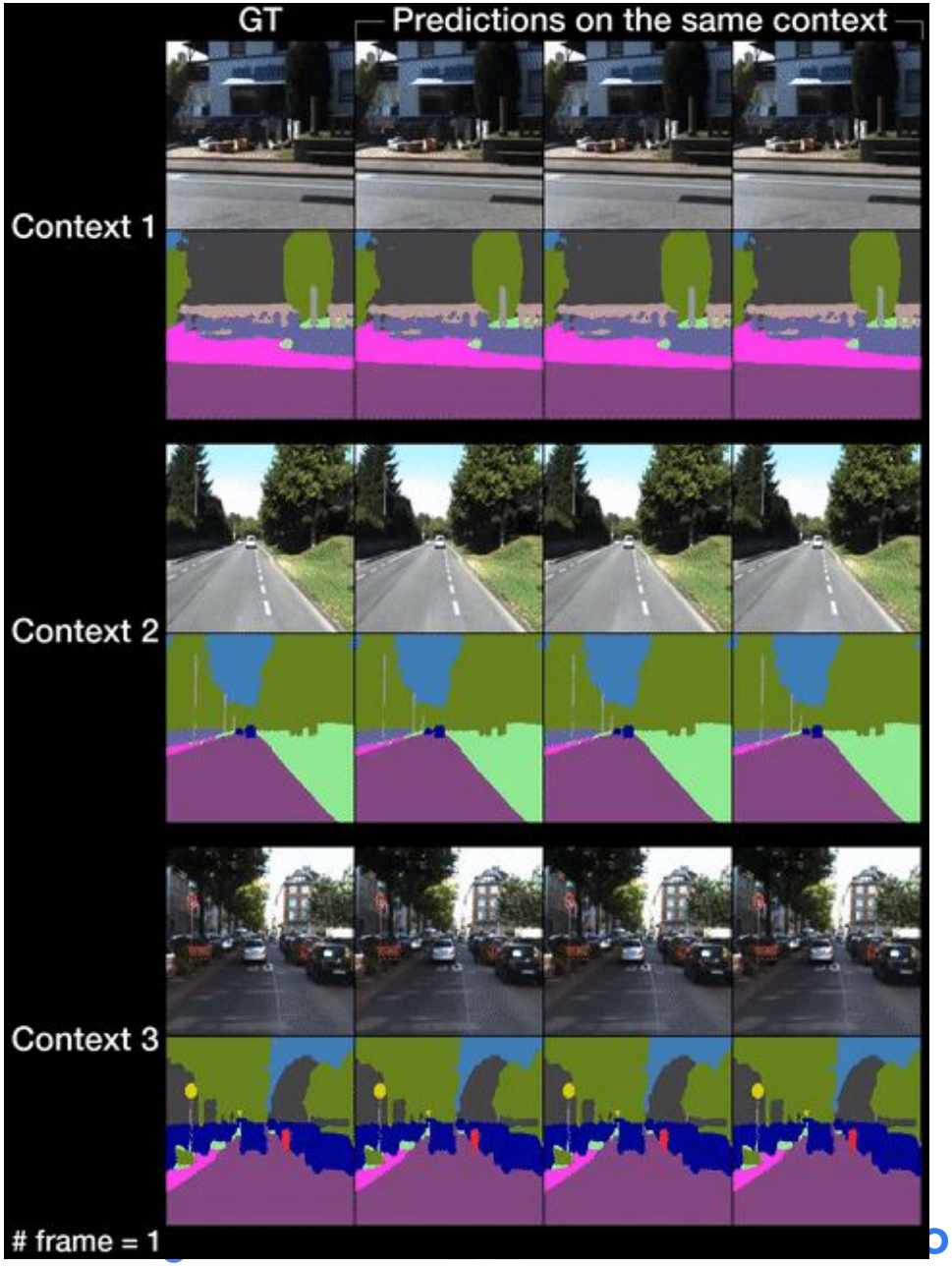
Model	t=1	t=250	t=400
SVG-Extend	13.1	22.9	27.2
Bayes-WD-SL	23.0	6.6	6.8
Ours	**63.9**	**70.5**	**66.0**

Human evaluation (most-preferred ratio)

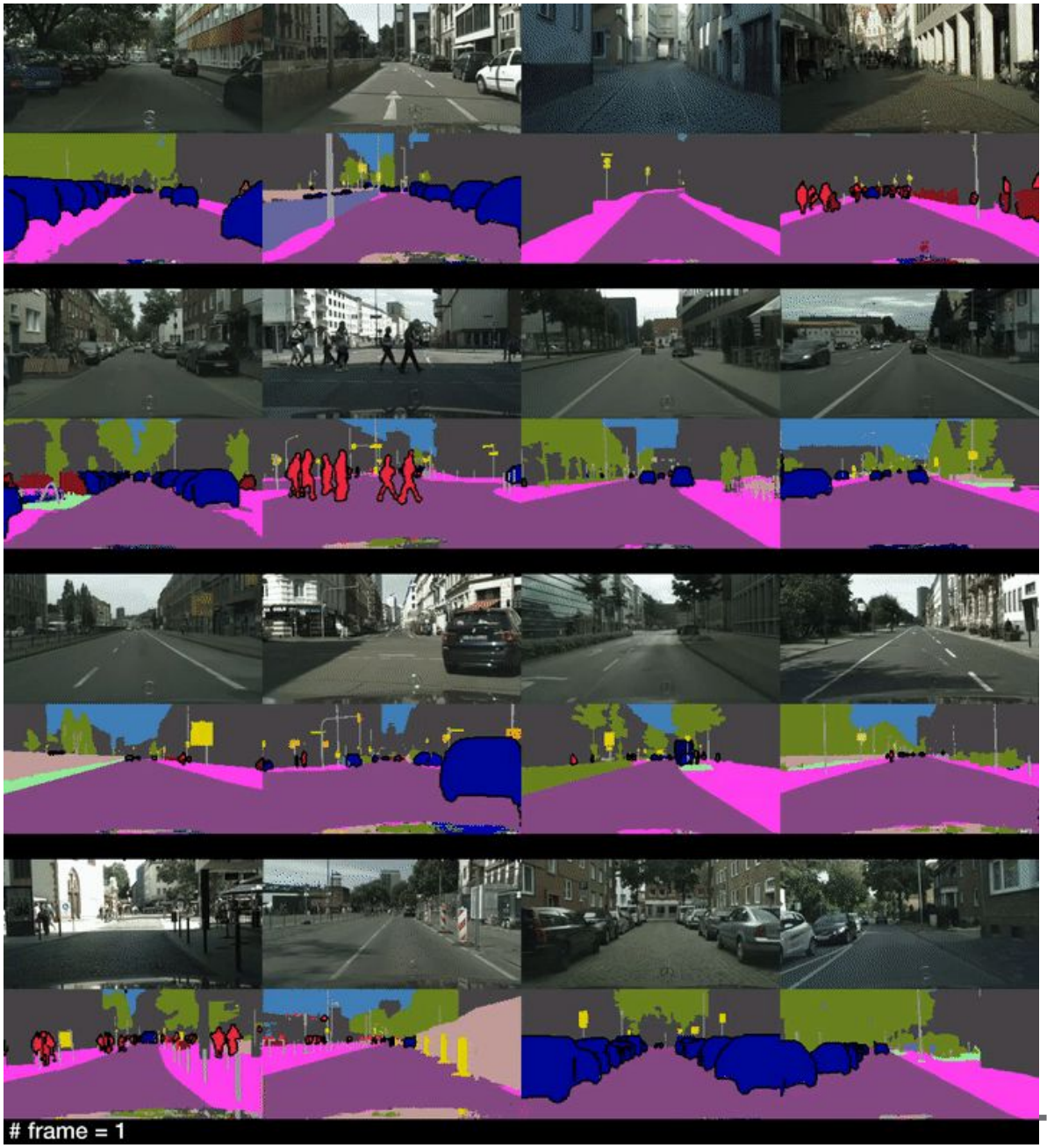
Results: Scaling up to high-resolution (256x256)



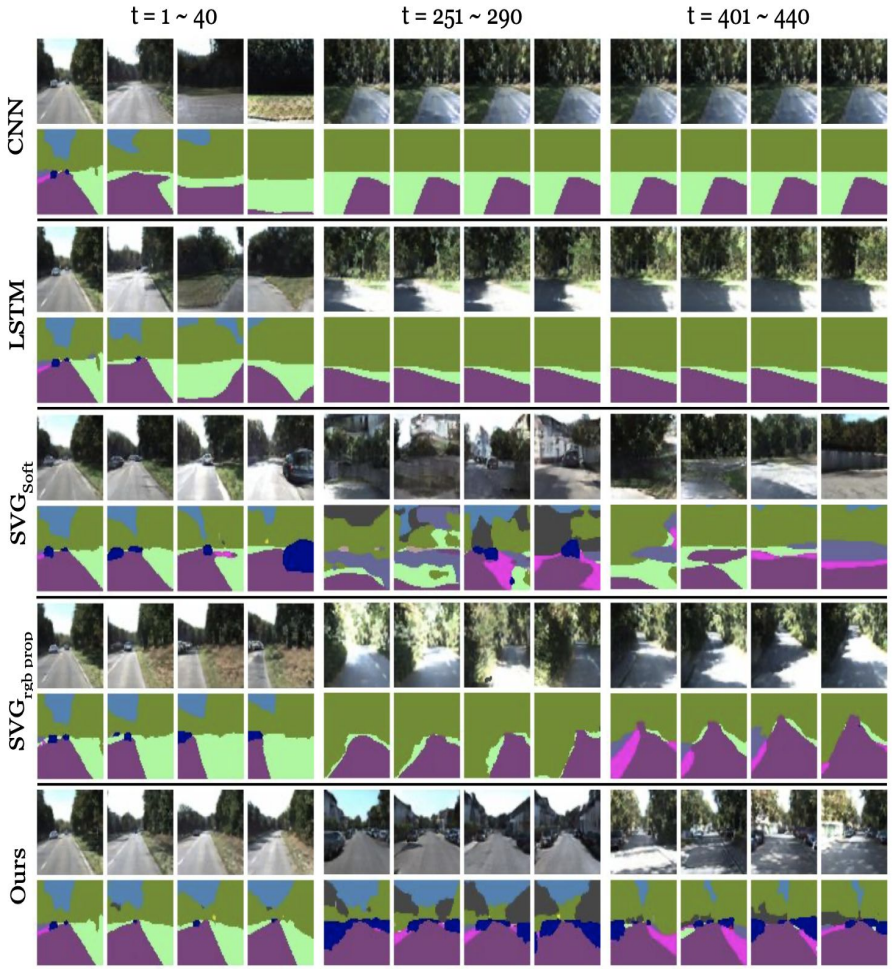
Results: Diverse predictions (256x256)



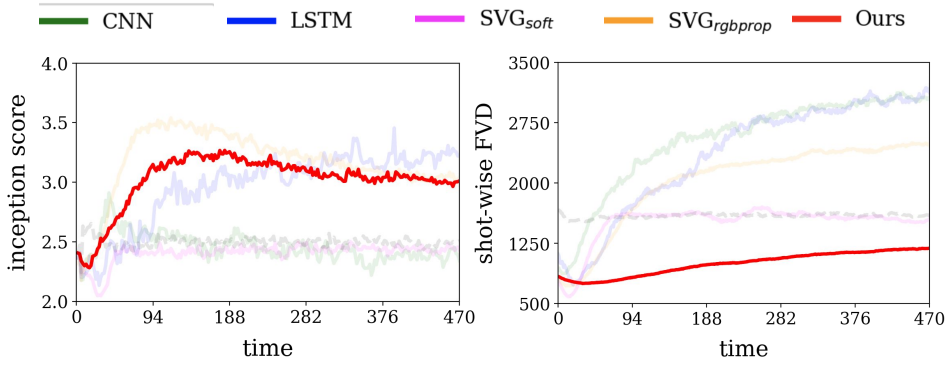
Results: Scaling up to high-resolution (256x512)



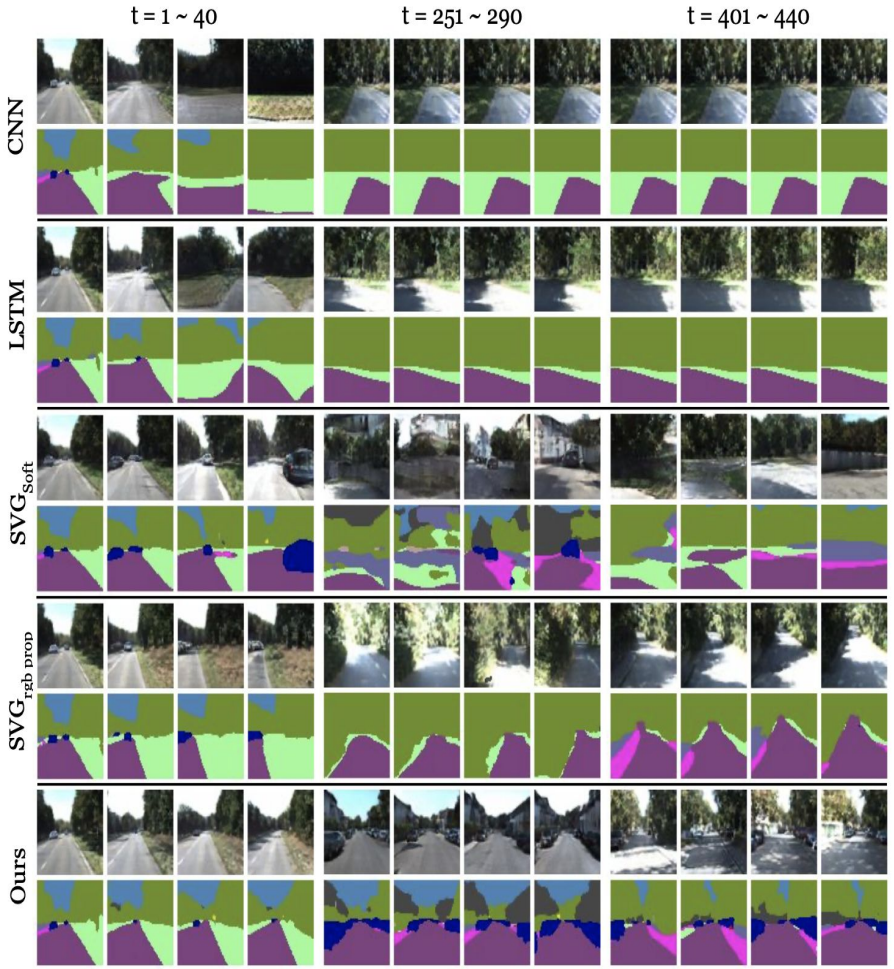
Results: Ablation Study (64x64)



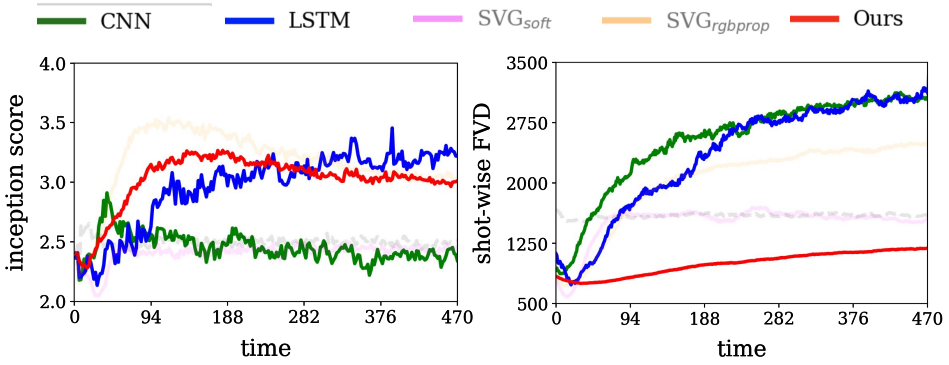
Method	Stochastic estimation	Recurrent estimation	Discretization	Independence to the image generator
LSTM	X	X	O	O
CNN	O	X	O	O
SVG_soft	O	O	X	O
SVG_rgbprop	O	O	O	X
Ours	**O**	**O**	**O**	**O**



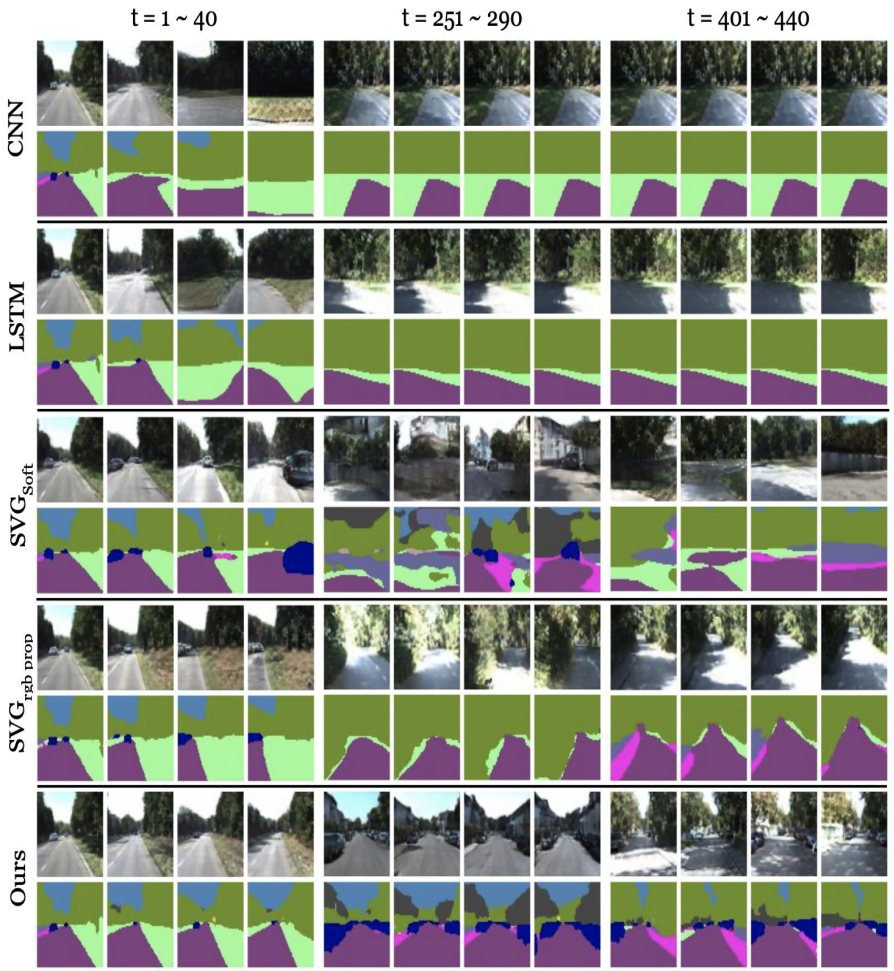
Results: Ablation Study (64x64)



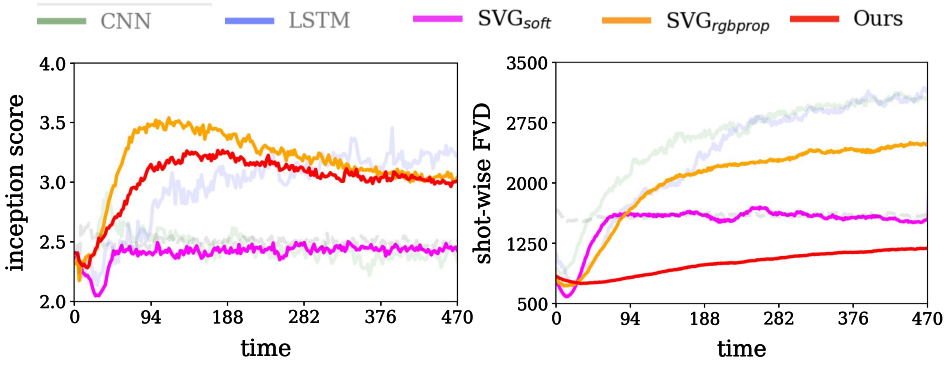
Method	Stochastic estimation	Recurrent estimation	Discretization	Independence to the image generator
LSTM	X	X	O	O
CNN	O	X	O	O
SVG_soft	O	O	X	O
SVG_rgbprop	O	O	O	X
Ours	O	O	O	O



Results: Ablation Study (64x64)



Method	Stochastic estimation	Recurrent estimation	Discretization	Independence to the image generator
LSTM	X	X	O	O
CNN	O	X	O	O
SVG_soft	O	O	X	O
SVG_rgbprop	O	O	O	X
Ours	O	O	O	O



Conclusion

- We propose the **hierarchical video prediction** model.
- Our method can synthesize the future of videos **an order of magnitude longer** than existing methods.

Full videos and codes are available at:
https://1konny.github.io/HVP/