VIDEO PREDICTION

GROUP MEMBERS

Nguyễn Như Quỳnh
Bùi Thị Thu Hương
Trương Thị Thu Thảo
Nguyễn Thị Thảo Nguyên

CONTRIBUTION

Member	Percentages	Detail of Contribution
Nguyễn Như Quỳnh (Leader)	35%	- Tìm Paper cả lần 1 và 2, Code và các thuật toán khác nhau. - Báo cáo Online - Chạy code SimVP (paper ban đầu) trên Moving MNIST, chạy code và fix lỗi code CrevNet (paper sau). - Làm toàn bộ Slide Report - Thuyết trình
Bùi Thị Thu Hương	25%	- Đọc Paper - Nhận chạy code SimVP (paper ban đầu) trên Moving MNIST, TaxiBJ, chạy trên KTH nhưng không hoàn thành nên chạy sang code CrevNet (paper sau – code do Quỳnh gửi).
Trương Thị Thu Thảo	20%	- Đọc Paper - Tìm hiểu các Dataset khác nhau
Nguyễn Thị Thảo Nguyên	20%	- Đọc Paper - Tìm hiểu các Dataset khác nhau

TABLE OF CONTENTS

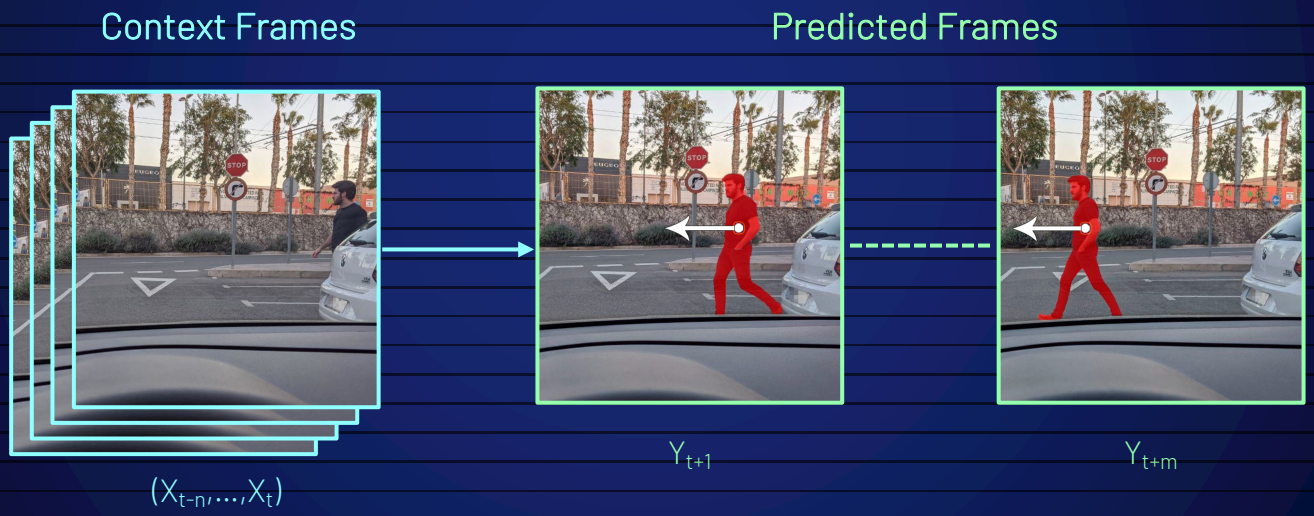
01

INTRODUCTION

BACKGROUND

Will the car hit the pedestrian?

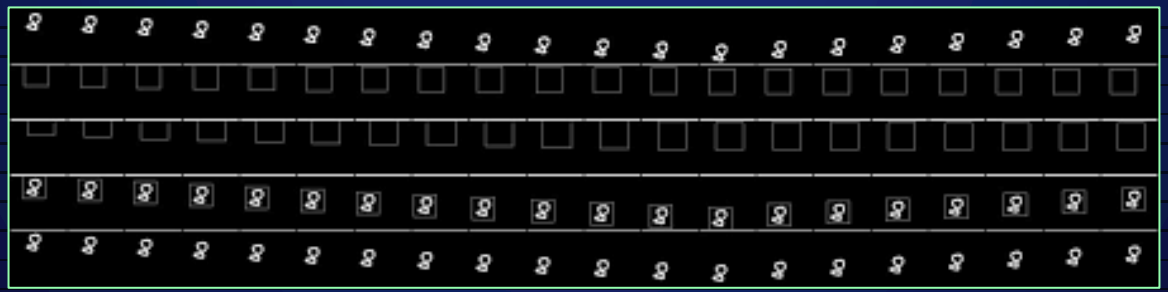
This example shows that predicting future frames in video is useful.



BACKGROUND

Video Prediction

- Video Prediction is the task of predicting future frames given past video frames.
- The ability to predict, anticipate and reason about future outcomes is a key component of intelligent decision-making systems. In light of the success of deep learning in computer vision, deep-learning-based video prediction emerged as a promising research direction.



DATASETS

Three datasets

Moving MNIST

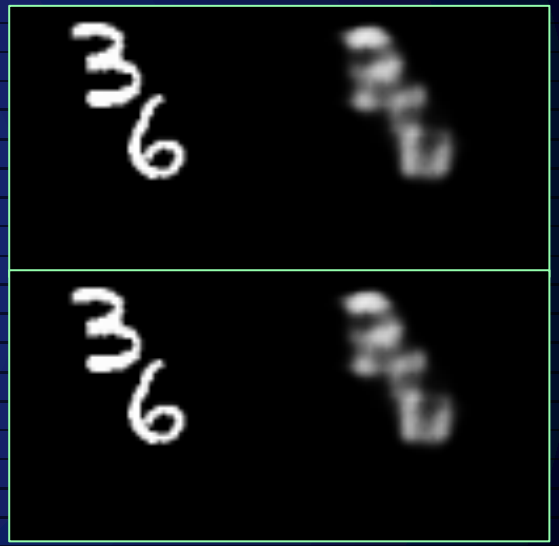
Traffic4cast

KITTI and Caltech Pedestrian

DATASETS

Moving MNIST (2015)

- Moving MNIST is a standard benchmark consisting of two digits independently moving within the 64 × 64 grid and bounced off the boundary.
- By assigning different initial locations and velocities to each digit, we can get an infinite number of sequences of length 20.
- Models are trained to predict the future 10 frames conditioned on the previous 10 frames.

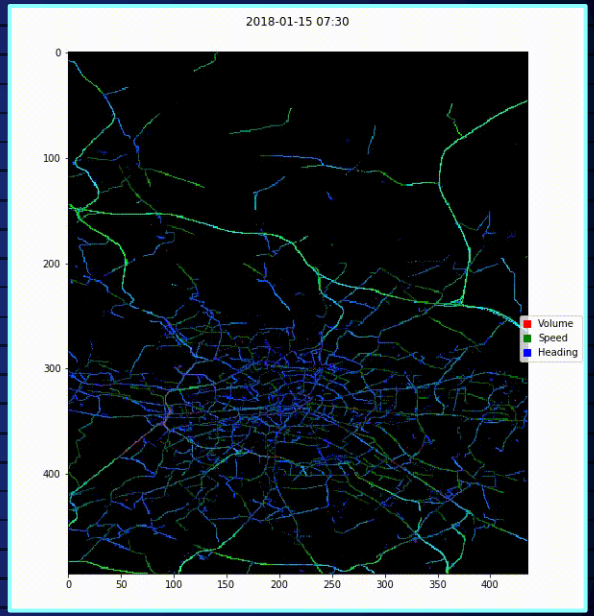


Demo of MMNIST

DATASETS

Traffic4cast (2019)

- The dataset collects the traffic statuses of 3 big cities over a year at a 5-minute interval.
- Each frame in dataset is a 495 × 436 × 3 heatmap, where the last dimension records 3 traffic statuses representing volume, mean speed and major direction at given location.
- We train each model to predict next 3 frames (the next 15 minutes) from 9 observations.



Demo of Traffic4cast

DATASETS

KITTI (2012) and Caltech Pedestrian (2009)

- KITTI and Caltect Pedestrian are car-mounted camera video datasets. We first center-crop all video frames and resize them into 128 × 160.

- Models are trained on KITTI dataset to predict the next frame after 10-frame warm-up and are evaluated on Caltech Pedestrian.

Demo of Caltech Pedestrian

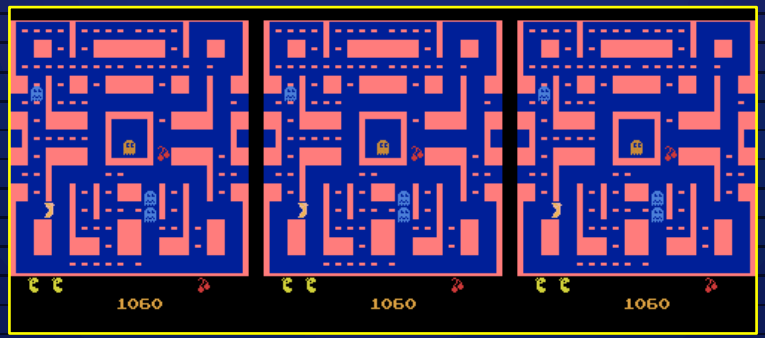
02

CREVNET MODEL

PROBLEM

Limitations of Classical Video Prediction

- Firstly, models are required to make pixel-wise predictions, which emphasizes the demand for the preservation of information through layers.
- Prior works attempt to address such demand through the extensive use of resolution-preserving blocks. Nevertheless, these resolution-preserving blocks are not guaranteed to preserve all the relevant information, and they greatly increase the memory consumption and computational cost of the models.



PROBLEM

Limitations of Classical Video Prediction

- The second drawback is that they cannot efficiently take advantage of 3D convolutions, as that would make these already cumbersome architectures even larger.
- 3D convolutions have been shown to be a very effective alternative to RNNs to capture temporal relations in a variety of video tasks (Liu et al. (2018); Carreira & Zisserman (2017)).

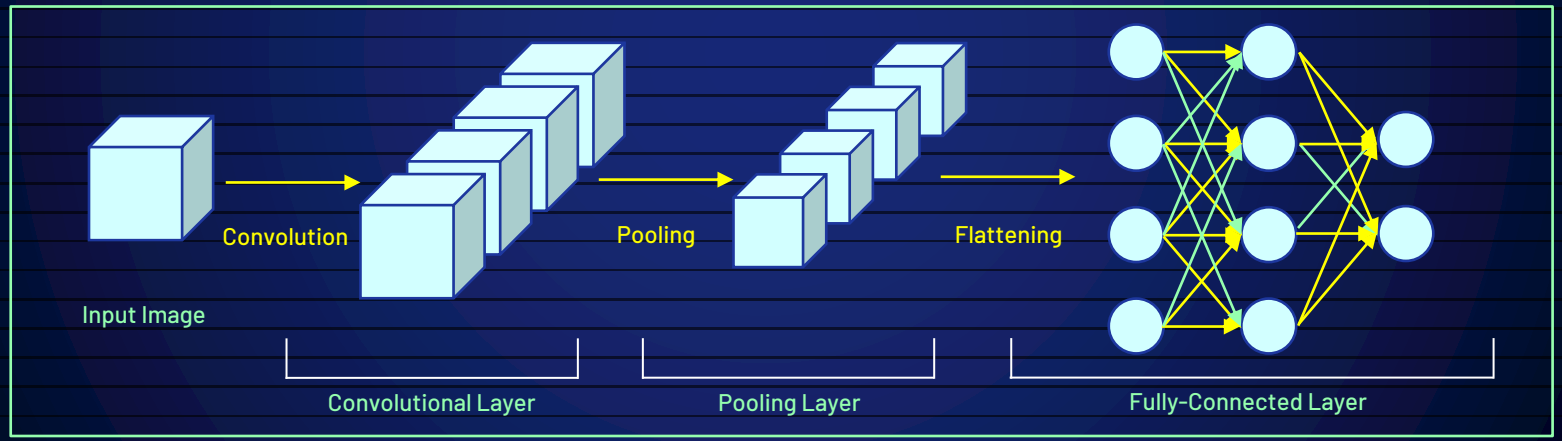


Figure. Basic 3D CNN Architecture

CREVNET MODEL

The Pipeline of CrevNet

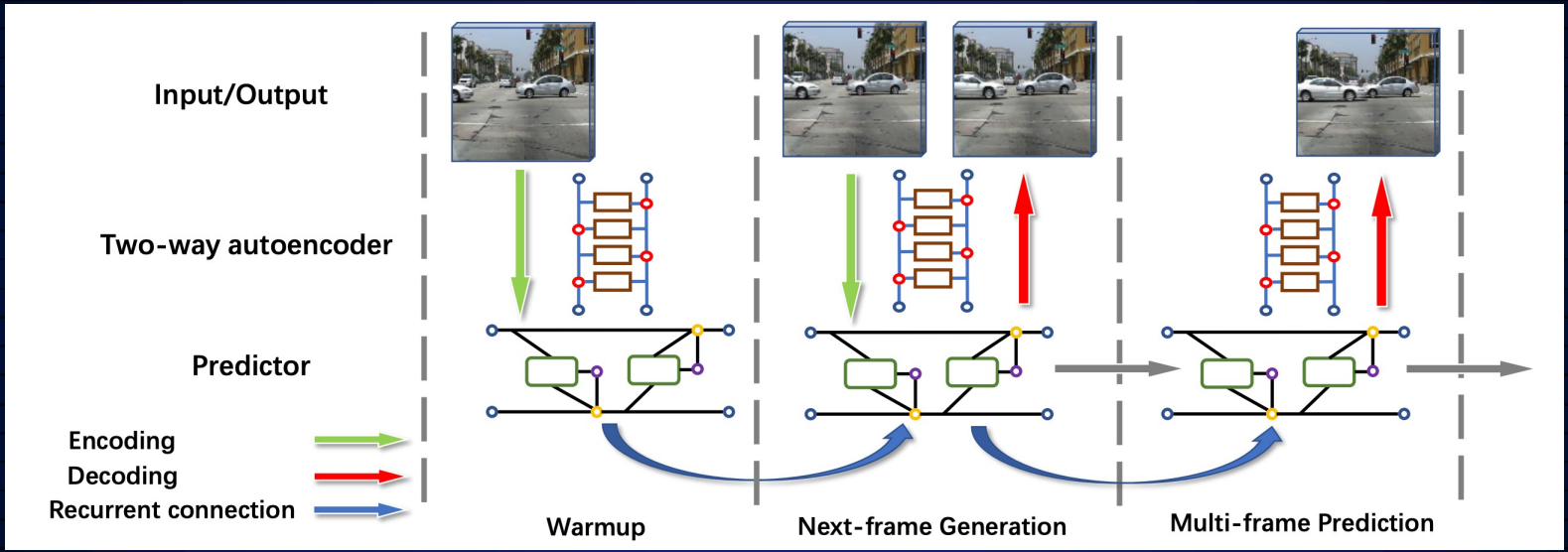


Figure. The pipeline of our proposed CrevNet where a single two-way autoencoder serves as both encoder and decoder. CrevNet first observes a warm-up video sequence and then starts a multi-frame video prediction without refeeding its own predictions.

CREVNET MODEL

The Network Architecture of CrevNet

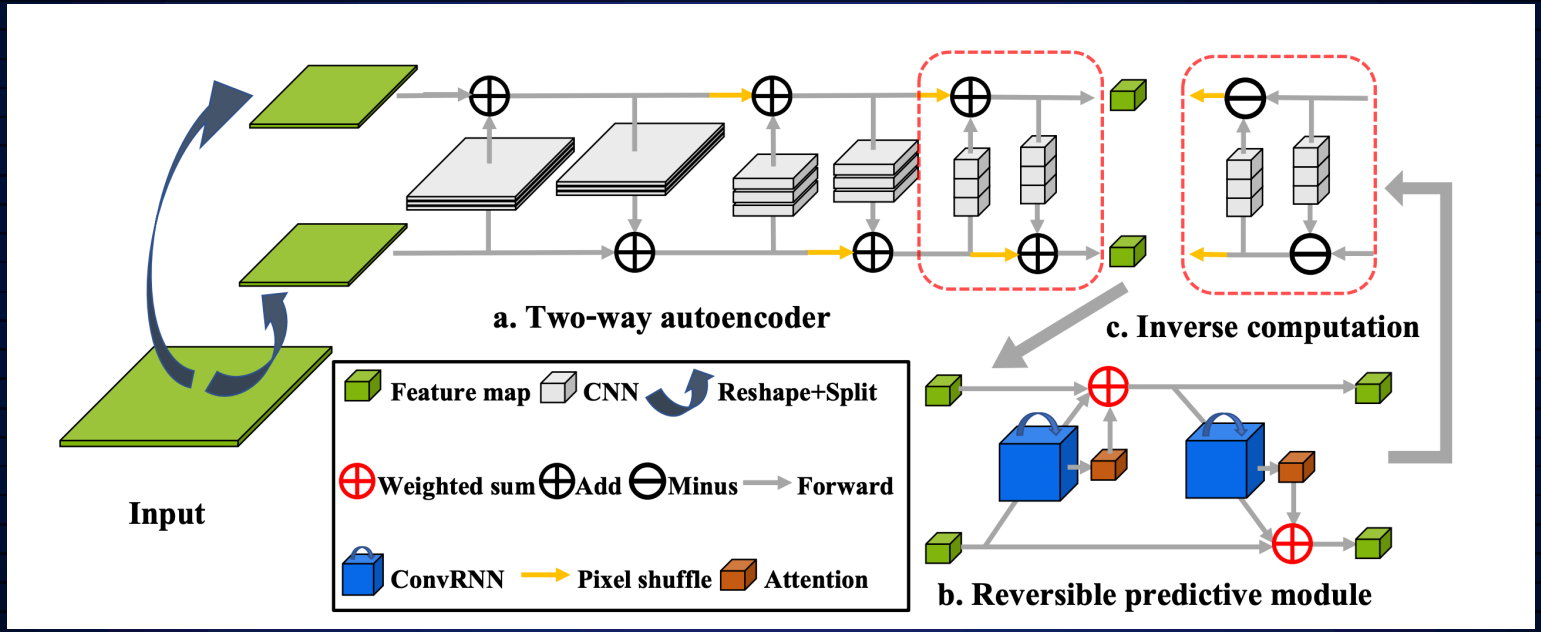


Figure. The network architecture of CrevNet

CREVNET MODEL

Reversible Predictive Module (RPM)

- Reversible predictive module (RPM) can be regarded as a recurrent extension of the two-way autoencoder. In the RPM, we substitute all standard convolutions with layers from the ConvRNN family (e.g. ConvLSTM or spatiotemporal LSTM) and introduce a soft attention (weighting gates) mechanism to form a weighted sum of the two groups instead of the direct addition.

- The main operations of RPM used in this paper are given as follows:

$$h_t^1 = \mathrm{ConvRNN}(x_t^1, h_{t-1}^1) \qquad \textit{ConvRNN}$$

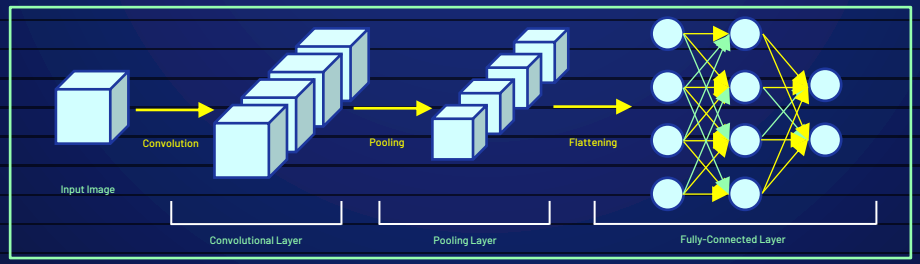
$$g_t = \phi(W_2 * \mathrm{ReLU}(W_1 * h_t^1 + b_1) + b_2) \qquad \textit{Attention module}$$

$$\hat{x_t^2} = (1 - g_t) \odot x_t^2 + g_t \odot h_t^1 \qquad \textit{Weighted sum}$$

CREVNET MODEL

3D Convolutions

- 3D convolutions are proposed to address the shortcomings of standard 2D convolutions. The major difference between 2D-CNNs and 3D-CNNs is that at each time step 2D-CNNs take as input one video frame, while 3D-CNNs read in and output a short video clip containing k continuous video frames.
- By applying convolutions on the temporal dimension along with the spatial dimension, models equipped with 3D convolution filters can not only extract representative spatiotemporal features, but also learn to produce consistent video clip at each generation, which further improve the quality of long-term prediction.



03

METRICS

METRICS

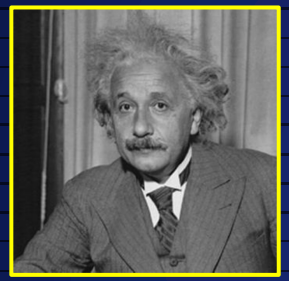
Mean Squared Error (MSE)

- MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.
- MSE is also commonly used in the training of video prediction models. The performance of each model in terms of per-frame MSE.

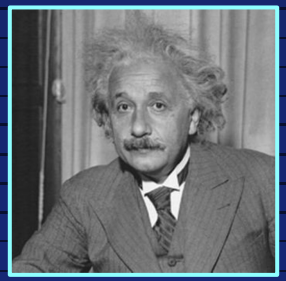
$$MSE = \frac{1}{N}\sum_{I=1}^{N}(y_i - \hat{y}_i)^2$$

METRICS

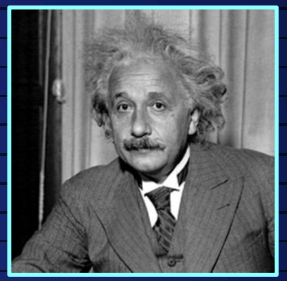
Structural Similarity Index Measure (SSIM)



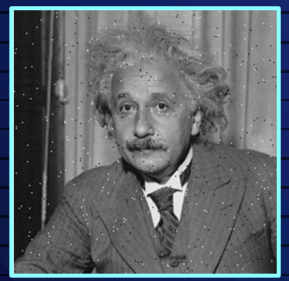
MSE = 0, SSIM = 1



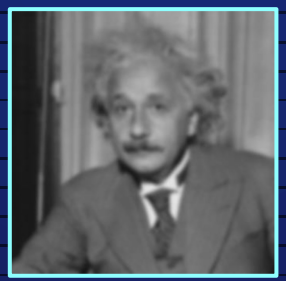
MSE = 144, SSIM = 0,988



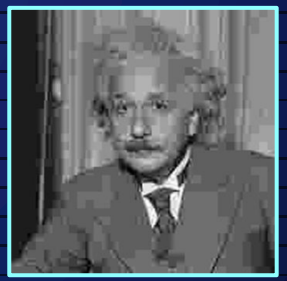
MSE = 144, SSIM = 0,913



MSE = 144, SSIM = 0,840



MSE = 144, SSIM = 0,694

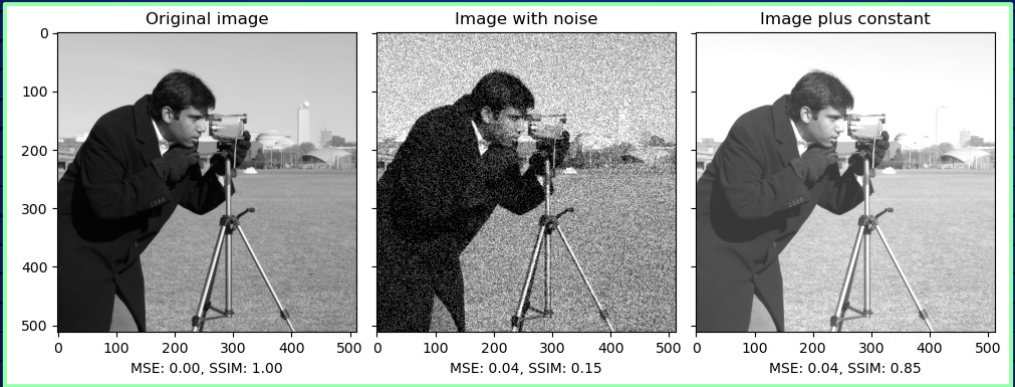


MSE = 142, SSIM = 0,662

METRICS

Structural Similarity Index Measure (SSIM)

- SSIM is a method for measuring the similarity between two images. The SSIM index can be viewed as a quality measure of one of the images being compared.
- In the example, all distorted images have roughly the same mean squared error (MSE) values with respect to the original image, but very different quality. SSIM gives a much better indication of image quality.



METRICS

Structural Similarity Index Measure (SSIM)

- The SSIM Index quality assessment index is based on the computation of three terms, namely the luminance term, the contrast term and the structural term. The overall index is a multiplicative combination of the three terms.

$$SSIM(x,y) = [l(x,y)]^{\alpha}.[c(x,y)]^{\beta}.[s(x,y)]^{\gamma}$$

$$where: \quad l(x,y) = \frac{2\mu_x\mu_y+C_1}{{\mu_x}^2+{\mu_y}^2+C_1}, c(x,y) = \frac{2\sigma_x\sigma_y+C_2}{{\sigma_x}^2+{\sigma_y}^2+C_2}, s(x,y) = \frac{\sigma_{xy}+C_3}{\sigma_x\sigma_y+C_3}$$

$\mu_x, \mu_y, \sigma_x, \sigma_y, \sigma_{xy}$ *are the local means, standard deviations and cross − covariance for image* x, y

04

RESULTS

RESULTS

Moving MNIST Dataset

- The general architecture of CrevNet used on Moving MNIST is composed of a 36-layer two-way autoencoder and 8 RPMs. All variants of CrevNet are trained by using the Adam optimizer with a starting learning rate of $5 \times 10{-}4$ to minimize MSE.

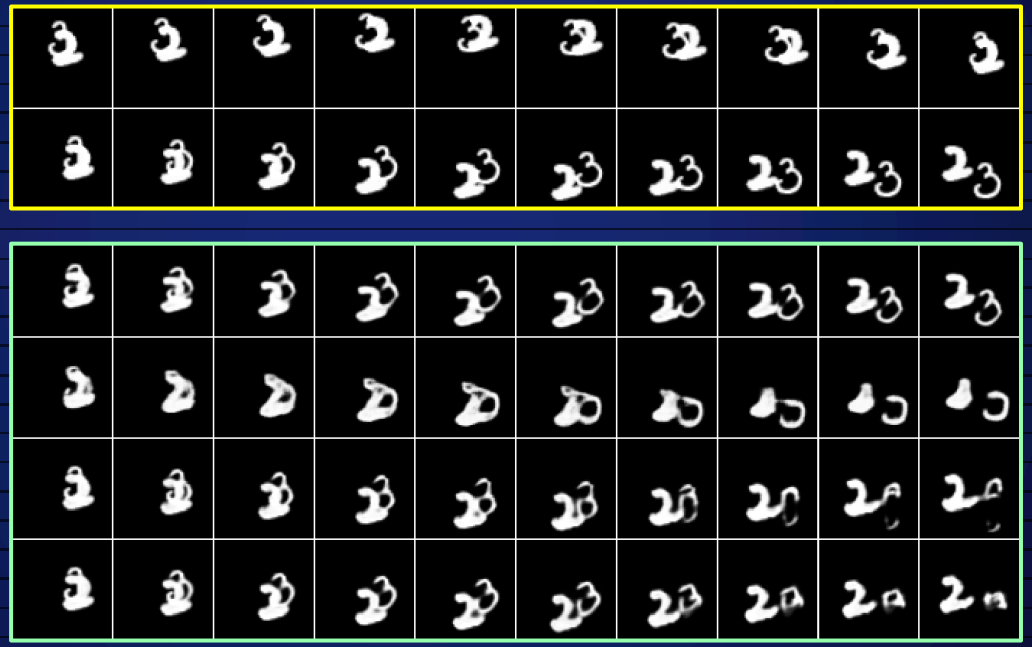
- The training process is stopped after 300 000 iterations with the batch size of 16 and evaluated with a fixed test set containing 5000 sequences.

Models	MSE	SSIM
PredRNN	56.8	0.869
PredRNN++	46.5	0.898
E3D-LSTM	41.3	0.910
CrevNet	22.3	0.949

Table 1. Quantitative evaluation of different methods on Moving MNIST

RESULTS

Moving MNIST Dataset



RESULTS

Traffic4cast Dataset

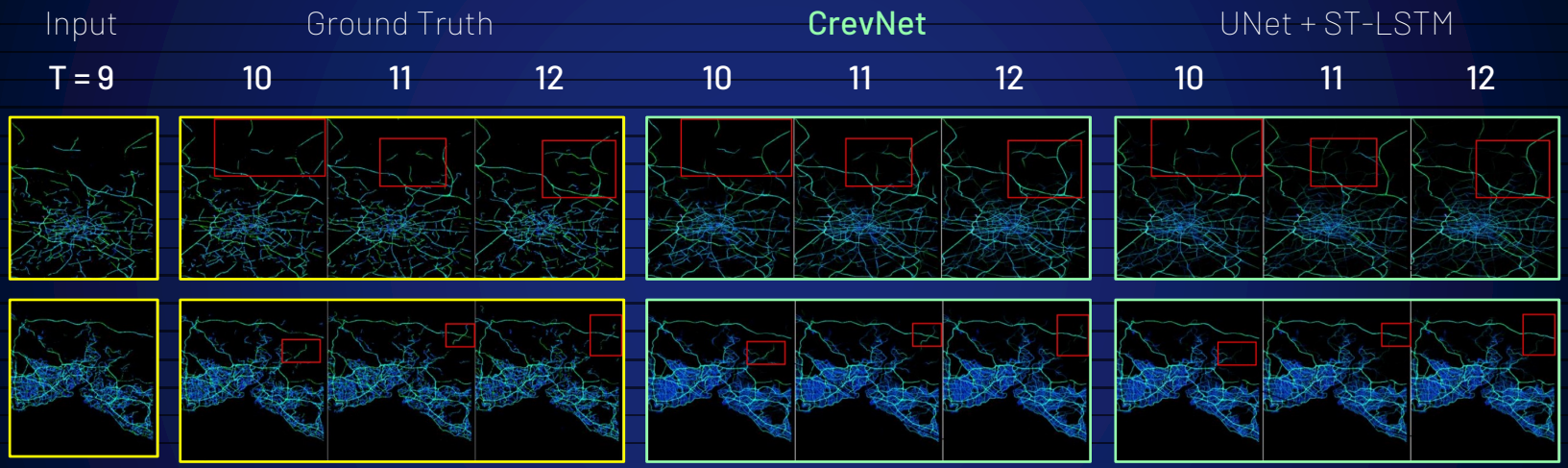


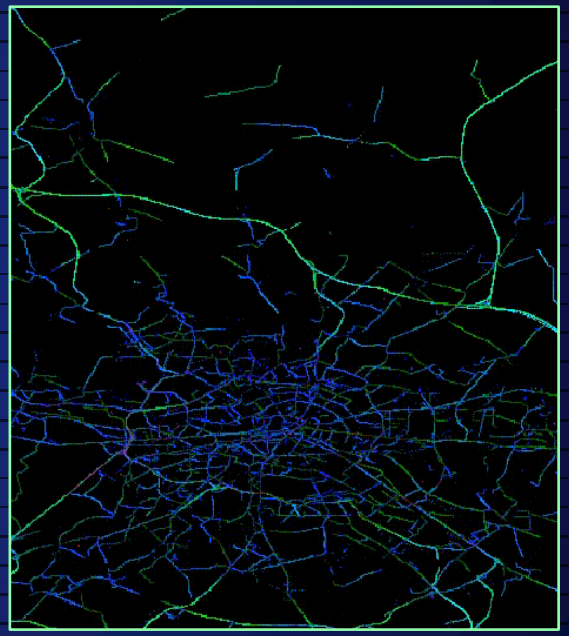
Figure. The visual comparison of Traffic4cast. The red boxes track some dynamics successfully captured by our CrevNet.

RESULTS

Traffic4cast Dataset

Table 2. Results on Traffic4cast Dataset

Models	MSE
UNet + ST-LSTM	9.725×10^{-3}
CrevNet	9.340×10^{-3}



RESULTS

KITTI and Caltech Pedestrian datasets

- Models are trained on KITTI dataset to predict the next frame after 10-frame warm-up and are evaluated on Caltech Pedestrian.
- The architecture of CrevNet used on KITTI is composed of a 48-layer two-way autoencoder.
- We add a 12-frame prediction comparison with CycleGAN (2019) and PredNet (2016).

Figure. The visual comparison of next-frame predictions on Caltech Pedestrian.

RESULTS

KITTI and Caltech Pedestrian datasets

Models	Next-Frame	3rd	6th	9th	12th	Average
			SSIM			
PredNet	0.905	0.72	0.66	0.61	0.58	0.701
CycleGAN	0.919	0.83	0.73	0.67	0.63	0.752
CrevNet	0.925	0.84	0.76	0.70	0.65	0.776

Table 3. Quantitative evaluation of different methods on Caltech Pedestrian dataset

RESULTS

KITTI and Caltech Pedestrian datasets

PredNet

CycleGAN

CrevNet

Ground Truth

Figure. The visual comparison of 12-frame prediction on Caltech Pedestrian. Notice how well CrevNet captures the detail and geometry of the buildings in the background, and the overall shading.

CONCLUSION

We described a novel conditionally reversible network, CrevNet, for pixel-level prediction of future frames in videos. The originality of our model lies in our use of the reversible two-way autoencoder and the accompanying reversible predictive module. Such architectural design enables the model to preserve fine-grained information without significant memory and computation overhead. CrevNet achieves state-of-the-art results on both synthetic and real-world datasets.

Thanks for listening!