7. Appendix

7.1. Dataset

Moving MNIST Moving MNIST [55] 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, thus enabling us to evaluate models without considering the data insufficiency issues. By default, models are trained to predict the future 10 frames conditioned on the previous 10 frames. Although the dynamics seem simple at first glance, accurately predicting consistent long-term future frames remains challenging for state-of-the-art models.

TrafficBJ TrafficBJ contains the trajectory data in Beijing collected from taxicab GPS with two channels, i.e. inflow or outflow defined in [75]. We choose data from the last four weeks as testing data, and all data before that as training data. Following the setting in [69], we transform the data into $[0, 1]$ via max-min normalization. Since the original data is between -1 and 1, the reported MSE and MAE are 1/4 and 1/2 of the original data respectively after transformation, which is consistent with previous literature [18,69].

Human3.6 Human3.6 [24] is a complex human pose dataset with 3.6 million samples, recording different activities such as taking photos, talking on the phone, posing, greeting, eating, etc. Similar to [18,55,69], only videos with "walking" scenario are used. Following previous works, we generate 4 future frames given the previous 4 RGB frames.

KITTI&Caltech Pedestrian KITTI [17] is one of the most popular datasets for mobile robotics and autonomous driving, as well as a benchmark for computer vision algorithms. It is composed by hours of traffic scenarios recorded with a variety of sensor modalities, including high-resolution RGB, gray-scale stereo cameras, and a 3D laser scanner. CalTech Pedestrian [11] is a driving dataset focused on detecting pedestrians. It is conformed of approximately 10 hours of 640×480 30 FPS video taken from a vehicle driving through regular traffic in an urban environment, making a total of 250,000 annotated frames distributed in 137 approximately minute-long segments. We follow the same protocol of PredNet [37] and CrevNet [73] for preprocessing, training and evaluation. Models are trained on KITTI dataset to predict the next frame after 10-frame warm-up and are evaluated on Caltech Pedestrian.

KTH The KTH dataset [53] contains 25 individuals performing 6 types of actions, i.e., walking, jogging, running, boxing, hand waving and hand clapping. Following [62,66], we use person 1-16 for training and 17-25 for testing. Models are trained to predict next 20 or 40 frames from the previous 10 observations.

7.2. Translator: should we use RNN, Transformer or CNN?

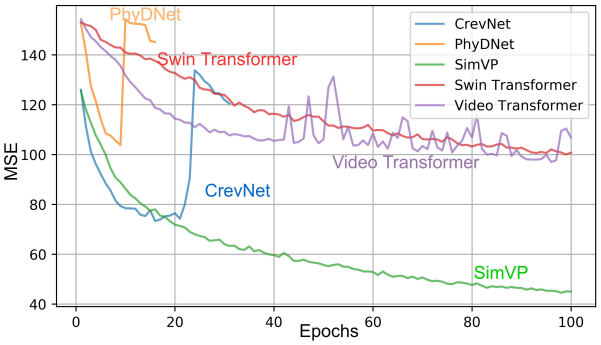

Figure 7. Training dynamics of various translators on Moving MNIST using large learning rate 0.01.

7.3. Network structure

Denote N_s, C_s as the layer numbers and hidden dimensions of the spatial Encoder (or Decoder). Similarly, N_t and C_t are the layer numbers and hidden dimensions of the Translator's encoder (or decoder). We use NNI (Neural Network Intelligence) to search hyperparameters, and the search space is shown in Table. 9. The final hyperparameter settings on various dataset can be found in Table. 10.

	Value
C_s	16,32,64
C_t	64,128,256,512
N_s	1,2,3,4
N_t	2,3,4,5

Table 9. Search space

	Human	MMNIST	TrafficBJ	Caltech	KTH
C_s	64	64	64	64	32
C_t	64	512	256	128	128
N_s	1	4	3	1	3
N_t	5	3	2	3	4

Table 10. Final hyperparameter setting.