

Machine Learning for Spatiotemporal Sequence Forecasting: A Survey

Xingjian Shi, Dit-Yan Yeung, *Senior Member, IEEE*

Abstract—Spatiotemporal systems are common in the real-world. Forecasting the multi-step future of these spatiotemporal systems based on the past observations, or, *Spatiotemporal Sequence Forecasting* (STSF), is a significant and challenging problem. Although lots of real-world problems can be viewed as STSF and many research works have proposed machine learning based methods for them, no existing work has summarized and compared these methods from a unified perspective. This survey aims to provide a systematic review of machine learning for STSF. In this survey, we define the STSF problem and classify it into three subcategories: *Trajectory Forecasting of Moving Point Cloud* (TF-MPC), *STSF on Regular Grid* (STSF-RG) and *STSF on Irregular Grid* (STSF-IG). We then introduce the two major challenges of STSF: 1) how to learn a model for multi-step forecasting and 2) how to adequately model the spatial and temporal structures. After that, we review the existing works for solving these challenges, including the general learning strategies for multi-step forecasting, the classical machine learning based methods for STSF, and the deep learning based methods for STSF. We also compare these methods and point out some potential research directions.

Index Terms—Spatiotemporal Sequence Forecasting, Machine Learning, Deep Learning, Data Mining.

1 INTRODUCTION

MANY real-world phenomena are spatiotemporal, such as the traffic flow, the diffusion of air pollutants and the regional rainfall. Correctly predicting the future of these spatiotemporal systems based on the past observations is essential for a wide range of scientific studies and real-life applications like traffic management, precipitation nowcasting, and typhoon alert systems.

If there exists an accurate numerical model of the dynamical system, which usually happen in the case that we have fully understood its rules, forecasting can be achieved by first finding the initial condition of the model and then simulate it to get the future predictions [1]. However, for some complex spatiotemporal dynamical systems like atmosphere, crowd, and natural videos, we are still not entirely clear about their inner mechanisms. In these situations, machine learning based methods, by which we can try to learn the systems' inner "laws" based on the historical data, have proven helpful for making accurate predictions [2], [3], [4], [5], [6]. Moreover, recent studies [7], [8] have shown that the "laws" inferred by machine learning

algorithms can further be used to guide other tasks like classification and control. For example, controlling a robotic arm will be much easier once you can anticipate how it will move after performing specific actions [8].

In this paper, we review these machine learning based methods for spatiotemporal sequence forecasting. We define a length- T spatiotemporal sequence as a series of matrices $\mathbf{X}_{1:T} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_T]$. Each $\mathbf{X}_t \in \mathbb{R}^{K \times (D+E)}$ contains a set of coordinates and their corresponding measurements. Here, K is the number of coordinates, D is the number of measurements, and E is the dimension of the coordinate. We further denote the measurement part and coordinate part of \mathbf{X}_t as $\mathbf{M}_t \in \mathbb{R}^{K \times D}$ and $\mathbf{C}_t \in \mathbb{R}^{K \times E}$. For many forecasting problems, we can use auxiliary information to enhance the prediction accuracy. For example, regarding wind speed forecasting, the latitudes and past temperatures are also helpful for predicting the future wind speeds. We denote the auxiliary information available at timestamp t as \mathcal{A}_t . The *Spatiotemporal Sequence Forecasting* (STSF) problem is to predict the length- L ($L > 1$) sequence in the future given the previous observations plus the auxiliary information that is allowed to be empty. The mathematical form is given in (1).

$$\hat{\mathbf{X}}_{t+1:t+L} = \arg \max_{\mathbf{X}_{t+1:t+L}} p(\mathbf{X}_{t+1:t+L} \mid \mathbf{X}_{1:t}, \mathcal{A}_t). \quad (1)$$

Here, we limit STSF to only contain problems where

- Xingjian Shi is with the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology. E-mail: xshiab@cse.ust.hk
- Dit-Yan Yeung is with the Department of Computer Science and Engineering, the Hong Kong University of Science and Technology. E-mail: dyeyung@cse.ust.hk

TABLE 1
Types of spatiotemporal sequence forecasting problems.

Problem Name	Coordinates	Measurements
Trajectory Forecasting of Moving Point Cloud	Changing	Fixed/Changing
Spatiotemporal Forecasting on Regular Grid	Fixed regular grid	Changing
Spatiotemporal Forecasting on Irregular Grid	Fixed irregular grid	Changing

the input and output are both sequences with multiple elements. According to this definition, problems like video generation using a static image [9] and dense optical flow prediction [10], in which the input or the output contain a single element, will not be counted.

Based on the characteristics of the coordinate sequence $\mathbf{C}_{1:T}$ and the measurement sequence $\mathbf{M}_{1:T}$, we classify STSF into three categories as shown in Table 1. If the coordinates are changing, the STSF problem is called *Trajectory Forecasting of Moving Point Cloud* (TF-MPC). Problems like human trajectory prediction [4], [11] and human dynamics prediction [12], [13], [14] fall into this category. We need to emphasize here that the entities in moving point cloud can not only be human but also be general moving objects. If the coordinates are fixed, we only need to predict the future measurements. Based on whether the coordinates lie in a regular grid, we can get the other two categories, called *STSF on Regular Grid* (STSF-RG) and *STSF on Irregular Grid* (STSF-IG). Problems like precipitation nowcasting [5], crowd density prediction [15] and video prediction [2], in which we have dense observations, are STSF-RG problems. Problems like air quality forecasting [3], influenza prediction [16] and traffic speed forecasting [17], in which the monitoring stations (or sensors) spread sparsely across the city, are STSF-IG problems. We need to emphasize that although the measurements in TF-MPC can also be time-variant due to factors like appearance change, most models for TF-MPC deal with the fixed-measurement case and only predict the coordinates. We thus focus on the fixed-measurement case in this survey.

Having both the characteristics of spatial data like images and temporal data like sentences and audios, spatiotemporal sequences contain information about “what”, “when”, and “where” and provide a comprehensive view of the underlying dynamical system. However, due to the complex spatial and temporal relationships within the data and the potential long forecasting horizon, spatiotemporal sequence forecasting imposes new challenges to the machine learning community. The first challenge is how to learn a model for multi-step forecasting. Early researches in time-series forecasting [18], [19] have investigated two approaches: *Direct Multi-step* (DMS) estimation and *Iterated Multi-step* (IMS) estimation. The DMS approach directly optimizes the multi-step forecasting

objective while the IMS approach learns a one-step-ahead forecaster and iteratively applies it to generate multi-step predictions. However, choosing between DMS and IMS involves a trade-off among forecasting bias, estimation variance, the length of the prediction horizon and the model’s nonlinearity [20]. Recent studies have tried to find the mid-ground between these two approaches [20], [21]. The second challenge is how to model the spatial and temporal structures within the data adequately. In fact, the number of free parameters of a length- T spatiotemporal sequence can be up to $O(K^T D^T)$ where K is usually above thousands for STSF-RG problems like video prediction. Forecasting in such a high dimensional space is impossible if the model cannot well capture the data’s underlying structure. Therefore, machine learning based methods for STSF, including both classical methods and deep learning based methods, have special designs in their model architectures for capturing these spatiotemporal correlations.

In this paper, we organize the literature about STSF following these two challenges. Section 2 covers the general learning strategies for multi-step forecasting including IMS, DMS, boosting strategy and scheduled sampling. Section 3 reviews the classical methods for STSF including the feature-based methods, state-space models, and Gaussian process models. Section 4 reviews the deep learning based methods for STSF including deep temporal generative models and feed-forward and recurrent neural network-based methods. Section 5 summarizes the survey and discusses some future research directions.

2 LEARNING STRATEGIES FOR MULTI-STEP FORECASTING

Compared with single-step forecasting, learning a model for multi-step forecasting is more challenging because the forecasting output $\hat{\mathbf{X}}_{t+1:t+L}$ is a sequence with non-i.i.d elements. In this section, we first introduce and compare two basic strategies called IMS and DMS [19] and then review two extensions that bridge the gap between these two strategies. To simplify the notation, we keep the following notation rules in this and the following sections. We denote the information available at timestamp t as $\mathcal{F}_t = \{\mathbf{X}_{1:t}, \mathcal{A}_t\}$. Learning an STSF model is thus to train a model, which can

be either probabilistic or deterministic, to make the L -step-ahead prediction $\hat{\mathbf{X}}_{t+1:t+L}$ based on \mathcal{F}_t . We also denote the ground-truth at timestamp t as $\tilde{\mathbf{X}}_t$ and recursively applying the basic function $f(x)$ for h times as $f^{(h)}(x)$. Setting the k th subscript to be “:” means to select all elements at the k th dimension. Setting the k th subscript to a letter i means to select the i th element at the k th dimension. For example, for matrix \mathbf{X} , $\mathbf{X}_{i,:}$ means the i th row, $\mathbf{X}_{:,j}$ means the j th column, and $\mathbf{X}_{i,j}$ means the (i, j) th element. The flattened version of matrices \mathbf{A} , \mathbf{B} and \mathbf{C} is denoted as $\text{vec}(\mathbf{A}, \mathbf{B}, \mathbf{C})$.

2.1 Iterative Multi-step Estimation

The IMS strategy trains a one-step-ahead forecasting model and iteratively feeds the generated samples to the one-step-ahead forecaster to get the multi-step-ahead prediction. There are two types of IMS models, the deterministic one-step-ahead forecasting model $\mathbf{X}_t = m(\mathcal{F}_t; \phi)$ and the probabilistic one-step-ahead forecasting model $p(\mathbf{X}_t | \mathcal{F}_t; \phi)$. For the deterministic forecasting model, the forecasting process and the training objective of the deterministic forecaster [20] are shown in (2) and (3) respectively where $d(\tilde{x}, x)$ measures the distance between \tilde{x} and x .

$$\hat{\mathbf{X}}_{t+h} = m^{(h)}(\mathcal{F}_t; \phi), \quad \forall 1 \leq h \leq L. \quad (2)$$

$$\phi^* = \arg \min_{\phi} \mathbb{E} \left[d(\tilde{\mathbf{X}}_{t+1}, m(\mathcal{F}_t; \phi)) \right]. \quad (3)$$

For the probabilistic forecasting model, we use the chain-rule to get the probability of the length- L prediction, which is given in (4), and estimate the parameters by *Maximum Likelihood Estimation* (MLE).

$$p(\mathbf{X}_{t+1:t+L} | \mathcal{F}_t; \phi) = p(\mathbf{X}_{t+1} | \mathcal{F}_t; \phi) \prod_{i=2}^L p(\mathbf{X}_{t+i} | \mathcal{F}_t, \mathbf{X}_{t+1:t+i-1}; \phi). \quad (4)$$

When the one-step-ahead forecaster is probabilistic and nonlinear, the MLE of (4) is generally intractable. In this case, we can draw samples from $p(\mathbf{X}_{t+1:t+L} | \mathcal{F}_t; \phi)$ by techniques like *Sequential Monte-Carlo* (SMC) [22] and predict based on these samples.

There are two advantages of the IMS approach: **1)** The one-step-ahead forecaster is easy to train because it only needs to consider the one-step-ahead forecasting error and **2)** we can generate predictions of an arbitrary length by recursively applying the basic forecaster. However, there is a discrepancy between training and testing in IMS. In the training process, we use the ground-truths of the previous $L-1$ steps to generate the L th-step prediction. While in the testing process, we feed back the model predictions instead of the ground-truths to the forecaster. This makes the model prone to accumulative errors in the forecasting process [21]. For

the deterministic forecaster example above, the optimal multi-step-ahead forecaster, which can be obtained by minimizing $\mathbb{E} \left[\sum_{h=1}^L d(\tilde{\mathbf{X}}_{t+h}, m^{(h)}(\mathcal{F}_t; \phi)) \right]$, is not the same as recursively applying the optimal one-step-ahead forecaster defined in (3) when m is nonlinear. This is because the forecasting error at earlier timestamps will propagate to later timestamps [23].

2.2 Direct Multi-step Estimation

The main motivation behind DMS is to avoid the error drifting problem in IMS by directly minimizing the long-term prediction error. Instead of training a single model, DMS trains a different model m_h for each forecasting horizon h . There can thus be L models in the DMS approach. When the forecasters are deterministic, the forecasting process and training objective are shown in (5) and (6) respectively where $d(\cdot, \cdot)$ is the distance measure.

$$\hat{\mathbf{X}}_{t+h} = m_h(\mathcal{F}_t; \phi_L), \quad \forall 1 \leq h \leq L. \quad (5)$$

$$\phi_1^*, \dots, \phi_L^* = \arg \min_{\phi_1, \dots, \phi_L} \mathbb{E} \left[d(\tilde{\mathbf{X}}_{t+1:t+L}, \hat{\mathbf{X}}_{t+1:t+L}) \right]. \quad (6)$$

$$\phi^* = \arg \min_{\phi} \mathbb{E} \left[\sum_{h=1}^L d(\tilde{\mathbf{X}}_{t+h}, m^{(h)}(\mathcal{F}_t; \phi)) \right]. \quad (7)$$

To disentangle the model size from the number of forecasting steps L , we can also construct m_h by recursively applying the one-step-ahead forecaster m_1 . In this case, the model parameters $\{\phi_1, \dots, \phi_L\}$ are shared and the optimization problem turns into (7). We need to emphasize here that (7), which optimizes the multi-step-ahead forecasting error, is different from (3), which only minimizes the one-step-ahead forecasting error. This construction method is widely adopted in deep learning based methods [7], [24].

2.3 Comparison of IMS and DMS

Chevillon [19] compared the IMS strategy and DMS strategy. According to the paper, DMS can lead to more accurate forecasts when 1) the model is misspecified, 2) the sequences are non-stationary, or 3) the training set is too small. However, DMS is more computationally expensive than IMS. For DMS, if the ϕ_h s are not shared, we need to store and train L models. If the ϕ_h s are shared, we need to recursively apply m_1 for $O(L)$ steps [19], [21], [25]. Both cases require larger memory storage or longer running time than solving the IMS objective. On the other hand, the training process of IMS is easy to parallelize since each forecasting horizon can be trained in isolation from the others [26]. Moreover, directly optimizing a L -step-ahead forecasting loss may fail when the forecasting model is highly nonlinear, or the parameters are not well-initialized [21].

Due to these trade-offs, models for STSF problems choose the strategies that best match the problems' characteristics. Generally speaking, DMS is preferred for short-term prediction while IMS is preferred for long-term forecast. For example, the IMS strategy is widely adopted in solving TF-MPC problems where the forecasting horizons are long, which are generally above 12 [4], [13] and can be as long as 100 [14]. The DMS strategy is used in STSF-RG problems where the dimensionality of data is usually very large and only short-term predictions are required [5], [6]. For STSF-IG problems, some works adopt DMS when the overall dimensionality of the forecasting sequence is acceptable for the direct approach [3], [27].

Overall, IMS is easier to train but less accurate for multi-step forecasting while DMS is more difficult to train but more accurate. Since IMS and DMS are somewhat complementary, later studies have tried to bridge the gap between these two approaches. We will introduce two representative strategies in the following two sections.

2.4 Boosting Strategy

The boosting strategy proposed in [20] combines the IMS and DMS for univariate time series forecasting. The strategy boosts the linear forecaster trained by IMS with several small and nonlinear adjustments trained by DMS. The model assumption is shown in (8).

$$\tilde{x}_{t+h} = m^{(h)}(\mathcal{F}_t; \phi) + \sum_{i=1}^{I_h} \nu l_i(\tilde{x}_{t-j}, \tilde{x}_{t-k}; \psi_i) + e_{t,h}. \quad (8)$$

Here, $m(\mathcal{F}_t; \phi)$ is the basic one-step-ahead forecasting model, $l_i(\tilde{x}_{t-j}, \tilde{x}_{t-k}; \psi_i)$ is the chosen weak learner at step i , I_h is the boosting iteration number, ν is the shrinkage factor and $e_{t,h}$ is the error term. The author set the base linear learner to be the auto-regressive model and set the weak nonlinear learners to be the penalized regression splines. Also, the weak learners were designed to only account for the interaction between two elements in the observation history. The training method is similar to the gradient boosting algorithm [28]. $m(\mathcal{F}_t; \phi)$ is first trained by minimizing the IMS objective (3). After that, the weak-learners are fit iteratively to the gradient residuals.

The author also performed a theoretical analysis of the bias and variance of different models trained with DMS, IMS and the boosting strategy for 2-step-ahead prediction. The result shows that the estimation bias of DMS can be arbitrarily small while that of the IMS will be amplified for larger forecasting horizon. The model estimated by the boosting strategy also has a bias, but it is much smaller than the IMS approach due to the compensation of the weak learners trained by DMS. Also, the variance of the DMS approach depends on the variability of the input history \mathcal{F}_t and the model

size which will be very large for complex and nonlinear models. The variance of the boosting strategy is smaller because the basic model is linear and the weak-learners only consider the interaction between two elements.

Although the author only proposed the boosting strategy for the univariate case, the method, which lies in the mid-ground between IMS and DMS, should apply to the STSF problems with multiple measurements, which leads to potential future works as discussed in Section 5.

2.5 Scheduled Sampling

The idea of *Scheduled Sampling* (SS) [21] is first to train the model with IMS and then gradually replaces the ground-truths in the objective function with samples generated by the model. When all ground-truth samples are replaced with model-generated samples, the training objective becomes the DMS objective. The generation process of SS is described in (9):

$$\begin{aligned} \forall 1 \leq h \leq L, \\ \hat{\mathbf{X}}_{t+h} &\sim p(\mathbf{X} \mid \mathcal{F}_t, \bar{\mathbf{X}}_{t+1:t+h-1}; \phi), \\ \bar{\mathbf{X}}_{t+h} &= (1 - \tau_{t+h})\hat{\mathbf{X}}_{t+h} + \tau_{t+h}\tilde{\mathbf{X}}_{t+h}, \\ \tau_{t+h} &\sim B(1, \epsilon_k). \end{aligned} \quad (9)$$

Here, $\hat{\mathbf{X}}_{t+h}$ and $\tilde{\mathbf{X}}_{t+h}$ are the generated sample and the ground-truth at forecasting horizon h , correspondingly. $p(\mathbf{X} \mid \mathcal{F}_t, \bar{\mathbf{X}}_{t+1:t+h-1}; \phi)$ is the basic one-step-ahead forecasting model. τ_{t+h} is a random variable following the binomial distribution, which controls whether to use the ground-truth or the generated sample. ϵ_k is the probability of choosing the ground-truth at the k th iteration. In the training phase, SS minimizes the distance between $\hat{\mathbf{X}}_{t+1:t+L}$ and $\tilde{\mathbf{X}}_{t+1:t+L}$, which is shown in (10). In the testing phase, τ is fixed to 0, meaning that the model-generated samples are always used.

$$\begin{aligned} &\mathbb{E}_{p(\tilde{\mathbf{X}}_{t+1:t+L} \mid \mathcal{F}_t; \phi)} [d(\hat{\mathbf{X}}_{t+1:t+L}, \tilde{\mathbf{X}}_{t+1:t+L})] \\ &= \mathbb{E}_{p(\tau_{t+1:t+L-1})} [d(\hat{\mathbf{X}}_{t+1:t+L}, \tilde{\mathbf{X}}_{t+1:t+L})] \\ &\quad \prod_{h=1}^L p(\hat{\mathbf{X}}_{t+h} \mid \mathcal{F}_t, \bar{\mathbf{X}}_{t+1:t+h-1}; \phi). \end{aligned} \quad (10)$$

If ϵ_k equals to 1, the ground-truth will always be chosen and the objective function will be the same as in the IMS strategy. If ϵ_k is 0, the generated sample will always be chosen and the optimization objective will be the same as in the DMS strategy. Based on this observation, the author proposed to gradually decay ϵ_k during training to make the optimization objective shift smoothly from IMS to DMS, which is a type of *curriculum learning* [29]. The paper also provided some ways to change the sampling ratio.

One issue of SS is that the expectation over τ_{t+h} in (10) makes the loss function non-differentiable. The

?

original paper [21] obviates the problem by treating \hat{X}_{t+h} s as constants, which does not optimize the true objective [25], [30].

The SS strategy is widely adopted in DL models for STSF. In [31], the author applied SS to video prediction and showed that the model trained via SS outperformed the model trained via IMS. In [24], the author proposed a training strategy similar to SS except that the training curriculum is adapted from IMS to DMS by controlling the number of the prediction steps instead of controlling the frequency of sampling the ground-truth. In [32], [33], the author used SS to train models for traffic speed forecasting and showed it outperformed IMS.

3 CLASSICAL METHODS FOR STSF

In this section, we will review the classical methods for STSF, including feature-based methods, state space models, and Gaussian process models. These methods are generally based on the shallow models or spatiotemporal kernels. Since a spatiotemporal sequence can be viewed as a multivariate sequence if we flatten the \mathbf{X}_t s into vectors or treat the observations at every location independently, methods designed for the general multivariate time-series forecasting are naturally applicable to STSF. However, directly doing so ignores the spatiotemporal structure within the data. Therefore, most algorithms introduced in this section have extended this basic approach by utilizing the domain knowledge of the specific task or considering the interactions among different measurements and locations.

3.1 Feature-based Methods

Feature-based methods solve the STSF by training a regression model based on some human-engineered spatiotemporal features. To reduce the problem size, most feature-based methods treat the measurements at the K locations to be independent given the extracted features. The high-level IMS and DMS models of the feature-based methods are given in (11) and (12) respectively.

$$\hat{\mathbf{X}}_{t+1,k} = f_{\text{IMS}}(\Psi_k(\mathcal{F}_t); \phi), \quad (11)$$

$$\hat{\mathbf{X}}_{t+1:t+L,k} = f_{\text{DMS}}(\Psi_k(\mathcal{F}_t); \phi). \quad (12)$$

Here, $\Psi_k(\mathcal{F}_t)$ represents the features at location k . Once the feature extraction method Ψ_k is defined, any regression model f can be applied. The feature-based methods are usually adopted in solving STSF-IG problems. Because the observations are sparsely and irregularly distributed in STSF-IG, it is sometimes more straightforward to extract human-engineered features than to develop a purely machine learning based model. Despite their straightforwardness, the feature-based methods have a major drawback. The model's

performance relies heavily on the quality of the human-engineered features. We will briefly introduce two representative feature-based methods for two STSF-IG problems to illustrate the approach.

Ohashi and Torgo [34] proposed a feature-based method for wind-speed forecasting. The author designed a set of features called spatiotemporal indicators and applied several regression algorithms including regression tree, support vector regression, and random forest. There are two kinds of spatiotemporal indicators defined in the paper: the average or weighted average of the observed historical values within a space-time neighborhood and the ratio of two average values calculated with different distance thresholds. The regression model takes the spatiotemporal indicators as the input to generate the one-step-ahead prediction. Experiment results show that the spatiotemporal indicators are helpful for enhancing the forecasting accuracy.

Zheng et al. [3] proposed a feature-based method for air quality forecasting. The author adopted the DMS approach and combined multiple predictors, including the temporal predictor, the spatial predictor, and the inflection predictor, to generate the features at location k and timestamp t . The temporal predictor focuses on the temporal side of the data and uses the measurements within a temporal sliding window to generate the features. The spatial predictor focuses on the spatial side of the data and uses the previous measurements within a nearby region to generate the features. Also, to deal with the sparsity and irregularity of the coordinates, the author partitioned the neighborhood region into several groups and aggregated the inner measurements within a group. The inflection predictor extracts the sudden changes within the observations. For the regression model, the temporal predictor uses the linear regression model, the spatial predictor uses a 2-hidden-layer neural network and the inflection predictor used the regression tree model. A regression tree model further combines these predictors.

3.2 State Space Models

状态空间模型

The *State Space Model* (SSM) adopts a generative viewpoint of the sequence. For the observed sequence $\mathbf{X}_{1:T}$, SSM assumes that each element \mathbf{X}_t is generated by a hidden state \mathbf{H}_t and these hidden states, which are allowed to be discrete in our definition, form a Markov process. The general form of the state space model [35] is given in (13). Here, g is the transition model, f is the observation model, \mathbf{H}_t is the hidden state, ϵ_t is the noise of the transition model, and σ_t is the noise of the observation model. The SSM has a long history in forecasting [36] and many classical time-series models like *Markov Model* (MM), *Hidden Markov Model* (HMM), *Vector Autoregression* (VAR),

and Autoregressive Integrated Moving Average Model (ARIMA) can be written as SSMs.

$$\begin{aligned}\mathbf{H}_t &= g(\mathbf{H}_{t-1}, \boldsymbol{\epsilon}_t), \\ \mathbf{X}_t &= f(\mathbf{H}_t, \boldsymbol{\sigma}_t).\end{aligned}\quad (13)$$

Given the underlying hidden state \mathbf{H}_t , the observation \mathbf{X}_t is independent concerning the historical information \mathcal{F}_t . Thus, the posterior distribution of the sequence for the next L steps can be written as follows:

$$\begin{aligned}p(\mathbf{X}_{t+1:t+L} | \mathcal{F}_t) &= \int p(\mathbf{H}_t | \mathcal{F}_t) \\ &\prod_{i=t+1}^{t+L} p(\mathbf{X}_i | \mathbf{H}_i) p(\mathbf{H}_i | \mathbf{H}_{i-1}) d\mathbf{H}_{t:t+L}\end{aligned}\quad (14)$$

The advantage of SSM for solving STSF is that it naturally models the uncertainty of the dynamical system due to its Bayesian formulation. The posterior distribution (14) encodes the probability of different future outcomes. We can not only get the most likely future but also know the confidence of our prediction by Bayesian inference. Common types of SSMs used in STSF literatures have either discrete hidden states or linear transition models and have limited representational power [37]. In this section, we will focus on these common types of SSMs and leave the review of more advanced SSMs, which are usually coupled with deep architectures, in Section 4.2. We will first briefly review the basics of three common classes of SSMs and then introduce their representative spatiotemporal extensions, including the group-based method, *Space-time Autoregressive Moving Average* (STARIMA), and the low-rank tensor auto-regression.

3.2.1 Common Types of SSMs

In STSF literature, the following three classes of SSMs are most common.

First-order Markov Model: When we set \mathbf{H}_t to be \mathbf{X}_{t-1} in (13) and constrain it to be discrete, the SSM reduces to the first-order MM. Assuming \mathbf{X}_t s have C choices, the first-order MM is determined by the transition matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$ where $\mathbf{A}_{i,j} = p(\mathbf{X}_{t+1} = i | \mathbf{X}_t = j)$ and the prior distribution $\pi \in \mathbb{R}^C$ where $\pi_i = p(\mathbf{X} = i)$. The maximum posterior estimator can be obtained via dynamic programming and the optimal model parameters can be solved in closed form [35].

Hidden Markov Model: When the states are discrete, the SSM turns into the HMM. Assuming the \mathbf{H}_t s have C choices, the HMM is determined by the transition matrix $\mathbf{A} \in \mathbb{R}^{C \times C}$, the state prior π and the class conditional density $p(\mathbf{X}_t | \mathbf{H}_t = i)$. For HMM, the maximum posterior estimator and the optimal model parameters can be calculated via the forward algorithm and the *Expectation Maximization*

(EM) algorithm, respectively [35]. It is worth emphasizing that the posterior of HMM is tractable even for non-Gaussian observation models. This is because the states of HMM are discrete and it is possible to enumerate all the possibilities.

Linear-Gaussian SSM: When the states are continuous and both g and f in (13) are linear-Gaussian, the SSM becomes the *Linear-Gaussian State Space Model* (LG-SSM) [35]:

$$\begin{aligned}\mathbf{H}_t &= \mathbf{A}_t \mathbf{H}_{t-1} + \boldsymbol{\epsilon}_t, \\ \mathbf{X}_t &= \mathbf{C}_t \mathbf{H}_t + \boldsymbol{\sigma}_t, \\ \boldsymbol{\epsilon}_t &\sim \mathcal{N}(0, \mathbf{Q}_t), \\ \boldsymbol{\sigma}_t &\sim \mathcal{N}(0, \mathbf{R}_t),\end{aligned}\quad (15)$$

where \mathbf{Q}_t and \mathbf{R}_t are covariance matrices of the Gaussian noises. A large number of classical time-series models like VAR and ARIMA can be written in the form of LG-SSM [35], [38]. As both the transition model and the observation model are linear-Gaussian, the posterior is also a linear-Gaussian and we can use the *Kalman Filtering* (KF) algorithm [39] to calculate the mean and variance. Similar to HMM, the model parameters can also be learned by EM.

3.2.2 Group-based Methods

Group-based methods divide the N locations into non-intersected groups and train an independent SSM for each group. To be more specific, for the location set $S = \{1, 2, \dots, N\}$, the group-based methods assume that it can be decomposed as $S = \cup_{i=1}^K G_i$ where $G_i \cap G_j = \emptyset, \forall i \neq j$. The measurement sequences that belong to the same group, i.e., $D_i = \{\{\mathbf{X}_{1,j}, \dots, \mathbf{X}_{t,j}\}, j \in G_i\}$, are used to fit a single SSM. If K is equal to 1, it reduces to the simplest case where the measurement sequences at all locations share the same model.

This technique has mainly been applied to TF-MPC problems. Asahara et al. [40] proposed a group-based extension of MM called *Mixed-Markov-chain Model* (MMM) for pedestrian-movement prediction. MMM assumes that the people belonging to a specific group satisfy the same first-order MM. To use the first-order MM, the author manually discretized the coordinates of each person. The group-assignments along with the model parameters are learned by EM. Experiments show that MMM outperforms MM and HMM in this task. Mathew et al. [41] extended the work by using HMM as the base model. Also, in [41], the groups are determined by off-line clustering algorithms instead of EM.

Group-based methods capture the intra-group relationships by model sharing. However, it cannot capture the inter-group correlations. Also, only a single group assignment is considered while there may be multiple grouping relationships in the spatiotemporal data.

Moreover, the latent state transitions of SSMs have some internal grouping effects. It is often not necessary to explicitly divide the locations into different groups. Due to these limitations, group-based methods are less mentioned in later research works.

3.2.3 STARIMA

STARIMA [42], [43], [44] is a classical spatiotemporal extension of the univariate ARIMA model [45]. STARIMA emphasizes the impact of a location's spatial neighborhoods on its future values. The fundamental assumption is that the future measurements at a specific location i is not only related to the past measurements observed at this location but also related to the past measurements observed at its local neighborhoods. STARIMA is designed for univariate STSF-RG and STSF-IG problems where there is a single measurement at each location. If we denote the measurements at timestamp t as $\mathbf{x}_t \in \mathbb{R}^N$, the model of STARIMA is shown in the following:

$$\begin{aligned} \Delta^d \mathbf{x}_t = & \sum_{k=1}^p \sum_{l=0}^{\lambda_k} \phi_{kl} \mathbf{W}^{(l)} \Delta^d \mathbf{x}_{t-k} + \epsilon_t \\ & - \sum_{k=1}^p \sum_{l=0}^{m_k} \theta_{kl} \mathbf{W}^{(l)} \epsilon_{t-k}. \end{aligned} \quad (16)$$

Here, Δ is the temporal difference operator where $\Delta \mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_{t-1}$ and $\Delta^k \mathbf{x}_t = \Delta\{\Delta^{k-1} \mathbf{x}_t\}$. ϕ_{kl} and θ_{kl} are respectively the autoregressive parameter and the moving average parameter at temporal lag k and spatial lag l . λ_k, m_k are maximum spatial orders. $\mathbf{W}^{(l)}$ is a predefined $N \times N$ matrix of spatial order l . ϵ_t is the normally distributed error vector at time t that satisfies the following relationships:

$$\begin{aligned} \mathbb{E}[\epsilon_t] &= \mathbf{0}, \\ \mathbb{E}[\epsilon_t \epsilon_{t+s}^T] &= \begin{cases} \sigma^2 \mathbf{I}, & s = 0 \\ \mathbf{0}, & \text{otherwise} \end{cases}, \\ \mathbb{E}[\mathbf{x}_t \epsilon_{t+s}^T] &= \mathbf{0}, \forall s > 0. \end{aligned} \quad (17)$$

The spatial order matrix $\mathbf{W}^{(l)}$ is the key in STARIMA for modeling the spatiotemporal correlations. $\mathbf{W}_{i,j}^{(l)}$ is only non-zero if the site j is the l th order neighborhood of site i . $\mathbf{W}^{(0)}$ is defined to be the identity matrix. The spatial order relationship in a two-dimensional coordinate system is illustrated in Figure 1. However, the $\mathbf{W}^{(l)}$ s need to be predefined by the model builder and cannot vary over-time. Also, the STARIMA still belongs to the class of LG-SSM. It is thus not suitable for modeling nonlinear spatiotemporal systems. Another problem of STARIMA is that

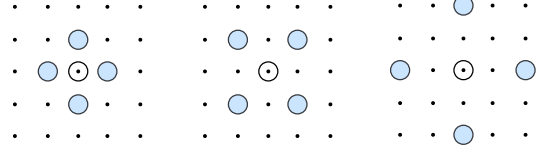


Fig. 1. Illustration of the spatial order relationships in STARIMA. From left to right are spatial order equals to 1, 2 and 3. The blue circles are neighbors of the white circle in the middle.

it is designed for problems where there is only one measurement at each location and has not considered the case of multiple correlated measurements.

3.2.4 Low-rank Tensor Auto-regression

Bahadori et al. [46] proposed to train an auto-regression model with low-rank constraint to better capture the latent structure of the spatiotemporal data. This method is designed for STSF-IG and STSF-RG problems where the coordinates are fixed. Assuming the observed measurements $\mathbf{X}_t \in \mathbb{R}^{N \times D}$ satisfy the VAR(J) model, the optimization problem of the low-rank tensor auto-regression is given in (18):

$$\begin{aligned} \min_{\mathcal{W}} \quad & \sum_{t=J+1}^T \|\hat{\mathbf{X}}_t - \mathbf{X}_t\|_F^2 + \mu \sum_{t=1}^T \text{tr}(\hat{\mathbf{X}}_t^T \mathbf{L} \hat{\mathbf{X}}_t), \\ \text{s.t.} \quad & r(\mathcal{W}) \leq \rho, \\ & \hat{\mathbf{X}}_{t+1,k} = \mathcal{W}_k \mathbf{X}_{t-J+1:t}, \forall t, k. \end{aligned} \quad (18)$$

Here, $\mathcal{W} \in \mathbb{R}^{N \times D \times JD}$ is the three-dimensional weight tensor. $\mathbf{L} \in \mathbb{R}^{D \times D}$ is the spatial Laplacian matrix calculated from the fixed coordinate matrix $\mathbf{C} \in \mathbb{R}^{N \times E}$. $r(\mathcal{W})$ means the rank of the weight tensor and ρ is the maximum rank. The regularization term in the objective function constrains the predicted results to be spatially smooth. μ is the strength of this spatial similarity constraint.

Constraining the rank of the weight tensor imposes an implicit spatiotemporal structure in the data. When the weight tensor is low-rank, the observed sequences can be described with a few latent factors.

The optimization problem (18) is non-convex and NP-hard in general [47] due to the low-rank constraint. One standard approach is to relax the rank constraint to a nuclear norm constraint and optimize the relaxed problem. However, this approach is not computationally feasible for solving (18) because of the high dimensionality of the spatiotemporal data. Thus, researchers [46], [48], [49] proposed optimization algorithms to accelerate the learning process.

The primary limitation of the low-rank tensor auto-regression is that the regression model is still linear even if we have perfectly learned the weights. We

cannot directly apply it to solve forecasting problems that have strong nonlinearities.

3.3 Gaussian Process

Gaussian Process (GP) is a non-parametric Bayesian model. GP defines a prior $p(f)$ in the function space and tries to infer the posterior $p(f \mid D)$ given the observed data [35]. GP assumes the joint distribution of the function values of an arbitrary finite set $\{x_1, x_2, \dots, x_n\}$, i.e., $p(f(x_1), \dots, f(x_n))$, is a Gaussian distribution. A typical form of the GP prior is shown in (19):

$$f \sim \mathcal{GP}(m(x), \kappa(x, x')), \quad (19)$$

where $m(x)$ is the mean function and $\kappa(x, x')$ is the kernel function.

In geostatistics, GP regression is also known as kriging [50], [51], which is used to interpolate the missing values of the data. When applied to STSF, GP is usually used in STSF-IG and STSF-RG problems where the coordinates are fixed. The general formula of the GP based models is shown in (20):

$$\begin{aligned} f &\sim \mathcal{GP}(m(X), \kappa_\theta), \\ Y &\sim p(Y \mid f(X)). \end{aligned} \quad (20)$$

Here, X contains the space-time positions of the samples, which is the concatenation of the spatial coordinates and the temporal index. For example, $X_k = (t_k, i_k, j_k)$ where t_k is the timestamp and (i_k, j_k) is the spatial coordinate. κ_θ is the kernel function parameterized by θ . $p(Y \mid f(X))$ is the observation model. Like SSM, GP is also a generative model. The difference is that GP directly generates the function describing the relationship between the observation and the future while SSM generates the future measurements one-by-one through the Markov assumption. We need to emphasize here that GP can also be applied to solve TF-MPC problems. A well-known example is the *Interacting Gaussian Processes* (IGP) [52], in which a GP models the movement of each person, and a potential function defined over the positions of different people models their interactions. Nevertheless, the method for calculating the posterior of IGP is similar to that of (20) and we will stick to this model formulation.

To predict the measurements given a GP model, we are required to calculate the predictive posterior $p(f^* \mid X^*, X^{\mathcal{D}}, Y^{\mathcal{D}})$ where $\{X^{\mathcal{D}}, Y^{\mathcal{D}}\}$ are the observed data points and X^* is the candidate space-time coordinates to infer the measurements. If the observation model is a Gaussian distribution, i.e., $p(Y \mid f(X)) = \mathcal{N}(f(X), \sigma^2)$, the posterior distribution is

also a Gaussian distribution as shown in (21).

$$\begin{aligned} p(f^* \mid X^*, X^{\mathcal{D}}, Y^{\mathcal{D}}) &= \mathcal{N}(\mu^*, \Sigma^*), \\ \mathbf{K} &= \kappa(X^{\mathcal{D}}, X^{\mathcal{D}}) + \sigma^2 \mathbf{I} \\ \mu^* &= \kappa(X^*, X^{\mathcal{D}}) \mathbf{K}^{-1} Y^{\mathcal{D}}, \\ \Sigma^* &= \kappa(X^*, X^*) - \kappa(X^*, X^{\mathcal{D}}) \mathbf{K}^{-1} \kappa(X^*, X^{\mathcal{D}}), \end{aligned} \quad (21)$$

If the observation model is non-Gaussian, which is common for STSF problems [53], Laplacian approximation can be utilized to approximate the posterior [35]. The inference process of GP is computationally expensive due to the involvement of \mathbf{K}^{-1} and the naive approach requires $O(|D|^3)$ computations and $O(|D|^2)$ storage. There are many acceleration techniques to speed up the inference [53]. We will not investigate them in detail in this survey and readers can refer to [51], [53].

The key ingredient of GP is the kernel function. For STSF problems, the kernel function should take both the spatial and temporal correlations into account. The common strategy for defining complex kernels is to combine multiple basic kernels by operations like summation and multiplication. Thus, GP models for STSF generally use different kernels to model different aspects of the spatiotemporal sequence and ensemble them together to form the final kernel function. In [53], the author proposed to decompose the spatiotemporal kernel as $\kappa_{st}((s, t), (s', t')) = \kappa_s(s, s') \kappa_t(t, t')$ where κ_s is the spatial kernel and κ_t is the temporal kernel [53]. If we denote the kernel matrix corresponding to $\kappa_{st}, \kappa_s, \kappa_t$ as $\mathbf{K}_{st}, \mathbf{K}_s, \mathbf{K}_t$, this decomposition results in $\mathbf{K}_{st} = \mathbf{K}_s \otimes \mathbf{K}_t$ where \otimes is the Kronecker product. Based on this observation, the author utilized the computational benefits of Kronecker algebra to scale up the inference of the GP models [54]. In [16], the author combined two spatial kernels and one temporal kernel to construct a GP model for seasonal influenza prediction. The author defined the basic spatial kernels as the average of multiple RBF kernels. The temporal kernel was defined to be the sum of the periodic kernel, the Pacioreks kernel, and the exponential kernel. The final kernel function used in the paper was $\kappa_{\text{final}} = \kappa_{\text{space}} + \kappa_{\text{time}} + \kappa_{\text{time}} \times \kappa_{\text{space2}}$.

3.4 Remarks

In this section, we reviewed three types of classical methods for solving the STSF problems, feature-based methods, SSMs, and GP models. The feature-based methods are simple to implement and can provide useful predictions in some practical situations. However, these methods rely heavily on the human-engineered features and may not be applied without the help of domain experts. The SSMs assume that the observations are generated by Markovian hidden states. We introduced three common types of SSMs: first-

order MM, HMM and LG-SSM. Group-based methods extend SSMs by creating groups among the states or observations. STARIMA extends ARIMA by explicitly modeling the impact of nearby locations on the future. Low-rank tensor autoregression extends the basic autoregression model by adding low-rank constraints on the model parameters. The advantage of these common types of SSMs for STSF is that we can calculate the exact posterior distribution and conduct Bayesian inference. However, the overall nonlinearity of these models are limited and may not be suitable for some complex STSF problems like video prediction. GP models the inner characteristics of the spatiotemporal sequences by spatiotemporal kernels. Although being very powerful and flexible, GP has high computational and storage complexity. Without any optimization, the inference of GP requires $O(|D|^3)$ computations and $O(|D|^2)$ storage, which makes it not suitable for cases when a huge number of training instances are available.

4 DEEP LEARNING METHODS FOR STSF

Deep Learning (DL) is an emerging new area in machine learning that studies the problem of how to learn a hierarchically-structured model to directly map the raw inputs to the desired outputs [26]. Usually, DL model stacks some basic learnable blocks, or layers, to form a deep architecture. The overall network is then trained end-to-end. In recent years, a growing number of researchers have proposed DL models for STSF. To better understand these models, we will start with an introduction of the relevant background of DL and then review the models for STSF.

4.1 Preliminaries

4.1.1 Basic Building Blocks

In this section, we will review the basic blocks that have been used to build DL models for STSF. Since there is a vast literature on DL, the introduction of this part serves only as a brief introduction. Readers can refer to [26] for more details.

Restricted Boltzmann Machine: *Restricted Boltzmann Machine* (RBM) [55] is the basic building block of *Deep Generative Models* (DGMs). RBM is an undirected graphical model that contains a layer of observable variables \mathbf{v} and another layer of hidden variables \mathbf{h} . These two layers are interconnected and form a bipartite graph. In its simplest form, \mathbf{h} and \mathbf{v} are all binary and contain n_h and n_v nodes. The

TABLE 2
Activations that appear in models reviewed in this survey.

Name	Formula
σ (sigmoid)	$\sigma(x) = \frac{1}{1+e^{-x}}$
tanh	$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$
ReLU [58]	$\text{ReLU}(x) = \max(0, x)$
Leaky ReLU [59]	$\text{LReLU}(x) = \max(-\alpha x, x)$

joint distribution and the conditional distributions are defined in (22) [26]:

$$\begin{aligned}
 p(\mathbf{v}, \mathbf{h}) &= \frac{1}{Z} \exp(-E(\mathbf{v}, \mathbf{h})), \\
 E(\mathbf{v}, \mathbf{h}) &= -\mathbf{b}^T \mathbf{v} - \mathbf{c}^T \mathbf{h} - \mathbf{v}^T \mathbf{W} \mathbf{h}, \\
 p(\mathbf{h} | \mathbf{v}) &= \prod_{j=1}^{n_h} \sigma \left((2\mathbf{h} - 1) \circ (\mathbf{c} + \mathbf{W}^T \mathbf{v}) \right)_j, \\
 p(\mathbf{v} | \mathbf{h}) &= \prod_{j=1}^{n_v} \sigma \left((2\mathbf{v} - 1) \circ (\mathbf{b} + \mathbf{W} \mathbf{h}) \right)_j.
 \end{aligned} \tag{22}$$

Here, \mathbf{b} is the bias of the visible state, \mathbf{c} is the bias of the hidden state, \mathbf{W} is the weight between visible states and hidden states, σ is the sigmoid function defined in Table 2, and Z is the normalization constant for energy-based models, which is also known as the partition function [35]. RBM can be learned by Contrastive Divergence (CD) [56], which approximates the gradient of $\log Z$ by drawing samples from k cycles of MCMC.

RBM
的训练方法

The basic form of RBM has been extended in various ways. *Gaussian-Bernoulli RBM* (GRBM) [57] assumes the visible variables are continuous and replaces the Bernoulli distribution with a Gaussian distribution. *Spike and Slab RBM* (ssRBM) maintains two types of hidden variables to model the covariance.

Activations: Activations refer to the element-wise nonlinear mappings $\mathbf{h} = f(\mathbf{x})$. Activation functions that are used in the reviewed models are listed in Table 2.

Sigmoid Belief Network: *Sigmoid Belief Network* (SBN) is another building block of DGM. Unlike RBM which is undirected, SBN is a directed graphical model. SBN contains a hidden binary state \mathbf{h} and a visible binary state \mathbf{v} . The basic generation process of SBN is given in the following:

$$\begin{aligned}
 p(\mathbf{v}_m | \mathbf{h}) &= \sigma(\mathbf{w}_m^T \mathbf{h} + \mathbf{c}_m), \\
 p(\mathbf{h}_j) &= \sigma(\mathbf{b}_j).
 \end{aligned} \tag{23}$$

Here, $\mathbf{b}_j, \mathbf{c}_m$ are biases and \mathbf{w}_m is the weight. Because the posterior distribution is intractable, SBN is learned by methods like *Neural Variational Inference and Learning* (NVIL) [60] and *Wake-Sleep* (WS) [61] algorithm, which train an inference network along with the generation network.

Fully-connected Layer: The *Fully-Connected* (FC) layer is the very basic building block of DL models. It refers to the linear-mapping from the input \mathbf{x} to the hidden state \mathbf{h} , i.e., $\mathbf{h} = \mathbf{W}\mathbf{x} + \mathbf{b}$, where \mathbf{W} is the weight and \mathbf{b} is the bias.

Convolution and Pooling: The convolution layer and pooling layer are first proposed to take advantage of the translation invariance property of image data. The convolution layer computes the output by scanning over the input and applying the same set of linear filters. Although the input can have an arbitrary dimensionality [62], we will just introduce 2D convolution and 2D pooling in this section. For input $\mathcal{X} \in \mathbb{R}^{C_i \times H_i \times W_i}$, the output of the convolution layer $\mathcal{H} \in \mathbb{R}^{C_o \times H_o \times W_o}$, which is also known as feature map, will be calculated as the following:

$$\mathcal{H} = \mathcal{W} * \mathcal{X} + \mathbf{b}. \quad (24)$$

Here, $\mathcal{W} \in \mathbb{R}^{C_i \times C_o \times K_h \times K_w}$ is the weight, K_h and K_w are the kernel sizes, “*” denotes convolution, and \mathbf{b} is the bias. If we denote the feature vector at the (i, j) position as $\mathcal{H}_{:,i,j}$, the flattened weight as $\mathbf{W} = \text{vec}(\mathcal{W})$, and the flattened version of the local region inputs as $\mathbf{x}_{\mathcal{N}(i,j)} = \text{vec}([\mathcal{X}_{:,s,t} \mid (s, t) \in \mathcal{N}(i, j)])$, (24) can be rewritten as (25):

$$\mathcal{H}_{:,i,j} = \mathbf{W}\mathbf{x}_{\mathcal{N}(i,j)} + \mathbf{b}. \quad (25)$$

This shows that convolution layer can also be viewed as applying multiple FC layers with shared weight and bias on local regions of the input. Here, the ordered neighborhood set $\mathcal{N}(i, j)$ is determined by hyper-parameters like the kernel size, stride, padding, and dilation [63]. For example, if the kernel size of the convolution filter is 3×3 , we have $\mathcal{N}(i, j) = \{(i + s, j + t) \mid -1 \leq s, t \leq 1\}$. Readers can refer to [26], [63] for details about how to calculate the neighborhood set $\mathcal{N}(i, j)$ based on these hyper-parameters. Also, the out-of-boundary regions of the input are usually set to zero, which is known as zero-padding.

Like convolution layer, the pooling layer also scans over the input and applies the same mapping function. Pooling layer generally has no parameter and uses some reduction operations like max, sum, and average for mapping. The formula of the pooling layer is given in the following:

$$\mathcal{H}_{k,i,j} = g(\{\mathcal{X}_{k,s,t} \mid (s, t) \in \mathcal{N}(i, j)\}), \quad (26)$$

where g is the element-wise reduction operation.

Deconvolution and Unpooling: The deconvolution [64] and unpooling layer [65] are first proposed to serve as the “inverse” operation of the convolution and the pooling layer. Computing the forward pass of these two layers is equivalent to computing the backward pass of the convolution and pooling layers [64].

Graph Convolution: Graph convolution generalizes the standard convolution, which is operated over a regular grid topology, to convolution over graph structures [33], [66], [67], [68]. There are two interpretations of graph convolution: the spectral interpretation and the spatial interpretation. In the spectral interpretation [66], [69], graph convolution is defined by the convolution theorem, i.e., the Fourier transform of convolution is the point-wise product of Fourier transforms:

$$\mathbf{x} *_G \mathbf{y} = F^{-1}(F(\mathbf{x}) \circ F(\mathbf{y})) = \mathbf{U}(\mathbf{U}^T \mathbf{x} \circ \mathbf{U}^T \mathbf{y}). \quad (27)$$

Here, $F(\mathbf{x}) = \mathbf{U}^T \mathbf{x}$ is the graph Fourier transform. The transformation matrix \mathbf{U} is defined based on the eigendecomposition of the graph Laplacian. Defferrard et al. [69] defined \mathbf{U} as the eigenvectors of $\mathbf{I} - \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$, in which \mathbf{D} is the diagonal degree matrix, \mathbf{A} is the adjacency matrix, and \mathbf{I} is the identity matrix.

In the spatial interpretation, graph convolution is defined by a localized parameter-sharing operator that aggregates the features of the neighboring nodes, which can also be termed as graph aggregator [33], [70]:

$$\mathbf{h}_i = \gamma_\theta(\mathbf{x}_i, \{\mathbf{z}_{\mathcal{N}_i}\}), \quad (28)$$

where \mathbf{h}_i is the output feature vector of node i , $\gamma_\theta(\cdot)$ is the mapping function parameterized by θ , \mathbf{x}_i is the input feature of node i , and $\mathbf{z}_{\mathcal{N}_i}$ contains the features of node i ’s neighborhoods. The mapping function needs to be permutation invariant and dynamically resizing. One example is the pooling aggregator defined in (29):

$$\mathbf{h}_i = \phi_o(\mathbf{x}_i \oplus \text{pool}_{j \in \mathcal{N}_i}(\phi_v(\mathbf{z}_j))), \quad (29)$$

where ϕ_o and ϕ_v are mapping functions and \oplus means concatenation. Compared with the spectral interpretation, the spatial interpretation does not require the expensive eigendecomposition and is more suitable for large graphs. Moreover, Defferrard et al. [69] and Kipf & Welling [68] proposed accelerated versions of the spectral graph convolution that could be interpreted from the spatial perspective.

4.1.2 Types of DL Models for STSF

Based on the characteristic of the model formulation, we can categorize DL models for STSF into two major types: *Deep Temporal Generative Models* (DTGMs) and *Feedforward Neural Network* (FNN) and *Recurrent Neural Network* (RNN) based methods. In the following, we will briefly introduce the basics about DTGM, FNN, and RNN.

DTGM: DTGMs are DGMs for temporal sequences. DTGM adopts a generative viewpoint of the sequence and tries to define the probability distribution $p(\mathbf{X}_{1:T})$. The aforementioned RBM and SBN are building-blocks of DTGM.

FNN: FNNs, also called deep feedforward networks [26], refer to deep learning models that construct the mapping $\mathbf{y} = f(\mathbf{x}; \theta)$ by stacking various of the aforementioned basic blocks such as FC layer, convolution layer, deconvolution layer, graph convolution layer, and activation layer. Common types of FNNs include *Multilayer Perceptron* (MLP) [26], [71], which stacks multiple FC layers and nonlinear activations, *Convolutional Neural Network* (CNN) [72], which stacks multiple convolution and pooling layers, and graph CNN [69], which stacks multiple graph convolution layers. The parameters of FNN are usually estimated by minimizing the loss function plus some regularization terms. Usually, the optimization problem is solved via stochastic gradient-based methods [26], in which the gradient is computed by backpropagation [73].

RNN: RNN generalizes the structure of FNN by allowing cyclic connections in the network. These recurrent connections make RNN especially suitable for modeling sequential data. Here, we will only introduce the discrete version of RNN which updates the hidden states using $\mathbf{h}_t = f(\mathbf{h}_{t-1}, \mathbf{x}_t; \theta)$ [26]. These hidden states can be further used to compute the output and define the loss function. One example of an RNN model is shown in (30). The model uses the tanh activation and the fully-connected layer for recurrent connections. It uses another full-connected layer to get the output:

$$\begin{aligned} \mathbf{h}_t &= \tanh(\mathbf{W}_h \mathbf{h}_{t-1} + \mathbf{W}_x \mathbf{x}_t + b), \\ \mathbf{o}_t &= \mathbf{W}_o \mathbf{h}_t + \mathbf{c}. \end{aligned} \quad (30)$$

Similar to FNNs, RNNs are also trained by stochastic gradient based methods. To calculate the gradient, we can first unfold an RNN to a FNN and then perform backpropagation on the unfolded computational graph. This algorithm is called *Backpropagation Through Time* (BPTT) [26].

Directly training the RNN shown in (30) with BPTT will cause the vanishing and exploding gradient problems [74]. One way to solve the vanishing gradient problem is to use the *Long-Short Term Memory* (LSTM) [75], which has a cell unit that is designed to capture the long-term dependencies. The formula of LSTM is given in the following:

$$\begin{aligned} \mathbf{i}_t &= \sigma(\mathbf{W}_{xi} \mathbf{x}_t + \mathbf{W}_{hi} \mathbf{h}_{t-1} + \mathbf{b}_i), \\ \mathbf{f}_t &= \sigma(\mathbf{W}_{xf} \mathbf{x}_t + \mathbf{W}_{hf} \mathbf{h}_{t-1} + \mathbf{b}_f), \\ \mathbf{c}_t &= \mathbf{f}_t \circ \mathbf{c}_{t-1} + \\ &\quad \mathbf{i}_t \circ \tanh(\mathbf{W}_{xc} \mathbf{x}_t + \mathbf{W}_{hc} \mathbf{h}_{t-1} + \mathbf{b}_c), \\ \mathbf{o}_t &= \sigma(\mathbf{W}_{xo} \mathbf{x}_t + \mathbf{W}_{ho} \mathbf{h}_{t-1} + \mathbf{b}_o), \\ \mathbf{h}_t &= \mathbf{o}_t \circ \tanh(\mathbf{c}_t). \end{aligned} \quad (31)$$

Here, \mathbf{c}_t is the memory cell and $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t$ are the input gate, the forget gate, and the output gate, respectively. The memory cell is accessed by the next layer when \mathbf{o}_t

is turned on. Also, the cell can be updated or cleared when \mathbf{i}_t or \mathbf{f}_t is activated. By using the memory cell and gates to control information flow, the gradient will be trapped in the cell and will not vanish over time, which is also known as the *Constant Error Carousel* (CEC) [75]. LSTM belongs to a broader category of RNNs called gated RNNs [26]. We will not introduce other architectures like *Gated Recurrent Unit* (GRU) in this survey because LSTM is the most representative example. To simplify the notation, we will also denote the transition rule of LSTM as $\mathbf{h}_t, \mathbf{c}_t = \text{LSTM}(\mathbf{h}_{t-1}, \mathbf{c}_{t-1}, \mathbf{x}_t; \mathbf{W})$.

4.2 Deep Temporal Generative Models

In this section, we will review DTGMs for STSF. Similar to SSM and GP model, DTGM also adopts a generative viewpoint of the sequences. The characteristics of DTGMs is that models belonging to this class can all be stacked to form a deep architecture, which enables them to model complex system dynamics. We will survey four DTGMs in the following. The general structures of these models are shown in Figure 2.

Temporal Restricted Boltzmann Machine: Sutskever & Hinton [76] proposed *Temporal Restricted Boltzmann Machine* (TRBM), which is a temporal extension of RBM. In TRBM, the bias of the RBM in the current timestamp is determined by the previous m RBMs. The conditional probability of TRBM is given in the following:

$$\begin{aligned} p(\mathbf{v}_t, \mathbf{h}_t \mid \hat{\mathbf{v}}_t, \hat{\mathbf{h}}_t) &= \frac{1}{Z} \exp \left(-E(\mathbf{v}_t, \mathbf{h}_t \mid \hat{\mathbf{v}}_t, \hat{\mathbf{h}}_t) \right), \\ E(\mathbf{v}_t, \mathbf{h}_t \mid \hat{\mathbf{v}}_t, \hat{\mathbf{h}}_t) &= -f_{b_v}(\hat{\mathbf{v}}_t)^T \mathbf{v}_t - f_{b_h}(\hat{\mathbf{v}}_t, \hat{\mathbf{h}}_t)^T \mathbf{h}_t \\ &\quad - \mathbf{v}_t^T \mathbf{W} \mathbf{h}_t, \\ f_{b_v}(\hat{\mathbf{v}}_t) &= \mathbf{a} + \sum_{i=1}^m \mathbf{A}_i^T \mathbf{v}_{t-i}, \\ f_{b_h}(\hat{\mathbf{v}}_t, \hat{\mathbf{h}}_t) &= \mathbf{b} + \sum_{i=1}^m \mathbf{B}_i^T \mathbf{v}_{t-i} + \sum_{i=1}^m \mathbf{C}_i^T \mathbf{h}_{t-i}, \end{aligned} \quad (32)$$

where $\hat{\mathbf{v}}_t = \mathbf{v}_{t-m:t-1}$ and $\hat{\mathbf{h}}_t = \mathbf{h}_{t-m:t-1}$. The joint distribution is modeled as the multiplication of the conditional probabilities, i.e., $p(\mathbf{v}_{1:T}, \mathbf{h}_{1:T}) = \prod_{i=1}^T p(\mathbf{v}_t, \mathbf{h}_t \mid \hat{\mathbf{v}}_t, \hat{\mathbf{h}}_t)$.

The inference of TRBM is difficult and involves evaluating the ratio of two partition functions [77]. The original paper adopted a heuristic inference procedure [76]. The results on a synthetic bouncing ball dataset show that a two-layer TRBM outperforms a single-layer TRBM in this task.

Recurrent Temporal Restricted Boltzmann Machine: Later, Sutskever et al. [77] proposed the *Recurrent Temporal Restricted Boltzmann Machine* (RTRBM) that extends the idea of TRBM by conditioning the parameters of the RBM at timestamp t on the

output of an RNN. The model of RTRBM is described in the following:

$$\begin{aligned}
 p(\mathbf{v}_t, \mathbf{h}_t | \mathbf{r}_{t-1}) &= \frac{1}{Z} \exp(-E(\mathbf{v}_t, \mathbf{h}_t | \mathbf{r}_{t-1})), \\
 E(\mathbf{v}_t, \mathbf{h}_t | \mathbf{r}_{t-1}) &= -\mathbf{h}_t^T \mathbf{W} \mathbf{v}_t - \mathbf{c}^T \mathbf{v}_t - \mathbf{b}^T \mathbf{h}_t \\
 &\quad - \mathbf{h}_t^T \mathbf{U} \mathbf{r}_{t-1}, \\
 \mathbf{r}_t &= \begin{cases} \sigma(\mathbf{W} \mathbf{v}_t + \mathbf{U} \mathbf{r}_{t-1} + \mathbf{b}), & t > 1 \\ \sigma(\mathbf{W} \mathbf{v}_t + \mathbf{b}_{init}), & t = 1 \end{cases}. \quad (33)
 \end{aligned}$$

Here, \mathbf{r}_t encodes the visible states $\mathbf{v}_{1:t}$. The inference is simpler in RTRBM than TRBM because the model is the same as the standard RBM once \mathbf{r}_{t-1} is given. Similar to RBM, RTRBM can also be learned by CD. The author also extended RTRBM to use GRBM for modeling continuous time-series. Qualitative experiments on the synthetic bouncing ball dataset show that RTRBM can generate more realistic sequences than TRBM.

Structured Recurrent Temporal Restricted Boltzmann Machine: Mittelman et al. [78] proposed *Structured Recurrent Temporal Restricted Boltzmann Machine* (SRTRBM) that improves upon RTRBM by learning the connection structure based on the spatiotemporal patterns of the input. SRTRBM uses two block-masking matrices to mask the weights between the hidden and visible nodes. The formula of SRTRBM is similar to RTRBM except for the usage of two masking matrices \mathbf{M}_W and \mathbf{M}_U :

$$\begin{aligned}
 E(\mathbf{v}_t, \mathbf{h}_t | \mathbf{r}_{t-1}) &= -\mathbf{h}_t^T \hat{\mathbf{W}} \mathbf{v}_t - \mathbf{c}^T \mathbf{v}_t \\
 &\quad - \mathbf{b}^T \mathbf{h}_t - \mathbf{h}_t^T \hat{\mathbf{U}} \mathbf{r}_{t-1}, \\
 \mathbf{r}_t &= \begin{cases} \sigma(\hat{\mathbf{W}} \mathbf{v}_t + \hat{\mathbf{U}} \mathbf{r}_{t-1} + \mathbf{b}), & t > 1 \\ \sigma(\hat{\mathbf{W}} \mathbf{v}_t + \mathbf{b}_{init}), & t = 1 \end{cases}, \quad (34) \\
 \hat{\mathbf{W}} &= \mathbf{W} \circ \mathbf{M}_W, \\
 \hat{\mathbf{U}} &= \mathbf{U} \circ \mathbf{M}_U.
 \end{aligned}$$

Here, \mathbf{M}_W and \mathbf{M}_U are block-masking matrices generated by a graph $G = (V; E)$. To construct the graph, the visible units and the hidden units are divided into $|V|$ nonintersecting subsets. Each node ϵ in the graph is assigned to a subset of hidden units C_ϵ^h and a subset of visible units C_ϵ^v . The ways to separate and assign the units are based on the spatiotemporal coordinates of the data. If there is an edge between two nodes in the graph, the corresponding hidden and visible subsets are connected in SRTRBM.

If the masks are fixed, the learning process of SRTRBM is the same as RTRBM. To learn the mask, the paper proposed to use the sigmoid function $\sigma(\nu_{ij})$ as a soft mask and learn the parameters ν_{ij} . Also, a sparsity regularizer is added to ν_{ij} s to encourage sparse connections. The training process is first to fix the mask while updating the other parameters and then

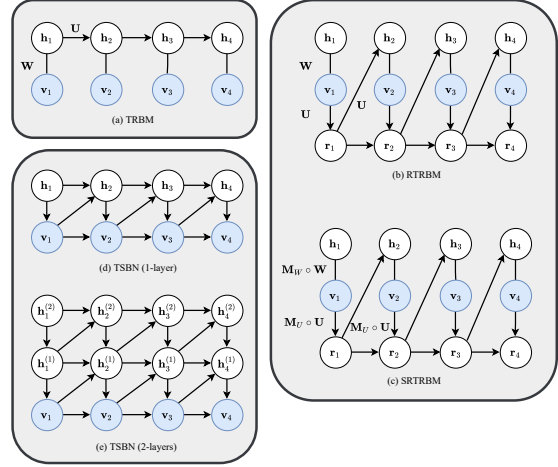


Fig. 2. Structure of the surveyed four types of deep temporal generative models: TRBM, RTRBM, SRTRBM and TSBN.

update the mask. The author also gives the ssRBM version of SRTBM and RTRBM by replacing RBM with ssRBM while fixing the part of the energy function related to \mathbf{r}_t to be unchanged. Experiments on human motion prediction, which is a TF-MPC task, and climate data prediction, which is an STSF-IG task, prove that SRTRBM achieves smaller one-step-ahead prediction error than RTRBM and RNN. This shows that using a masked weight matrix to model the spatial and temporal correlations is useful for STSF problems.

Temporal Sigmoid Belief Network: Gan et al. [79] proposed a temporal extension of SBN called *Temporal Sigmoid Belief Network* (TSBN). TSBN connects a sequence of SBNs in a way that the bias of the SBN at timestamp t depends on the state of the previous SBNs. For a length- T binary sequence $\mathbf{v}_{1:T}$, TSBN describes the following joint probability, in which we have omitted the biases for simplicity:

$$\begin{aligned}
 p_\theta(\mathbf{v}_{1:T}, \mathbf{h}_{1:T}) &= p(\mathbf{h}_1) p(\mathbf{v}_1 | \mathbf{h}_1) \\
 &\quad \prod_{t=2}^T p(\mathbf{h}_t | \mathbf{h}_{t-1}, \mathbf{v}_{t-1}) p(\mathbf{v}_t | \mathbf{h}_t, \mathbf{v}_{t-1}), \\
 p(\mathbf{h}_{t,j} = 1 | \mathbf{h}_{t-1}, \mathbf{v}_{t-1}) &= \sigma(\mathbf{A}_j^T \mathbf{h}_{t-1} + \mathbf{B}_j^T \mathbf{v}_{t-1}), \\
 p(\mathbf{v}_{t,j} = 1 | \mathbf{h}_t, \mathbf{v}_{t-1}) &= \sigma(\mathbf{C}_j^T \mathbf{h}_t + \mathbf{D}_j^T \mathbf{v}_{t-1}). \quad (35)
 \end{aligned}$$

Also, the author extended TSBN to the continuous domain by using a conditional Gaussian for the visible nodes:

$$\begin{aligned}
 p(\mathbf{v}_t | \mathbf{h}_t, \mathbf{v}_{t-1}) &= \mathcal{N}(\boldsymbol{\mu}_t, \text{diag}(\boldsymbol{\sigma}_t^2)), \\
 \boldsymbol{\mu}_t &= \mathbf{A}^T \mathbf{h}_{t-1} + \mathbf{B}^T \mathbf{v}_{t-1} + \mathbf{e}, \quad (36) \\
 \log(\boldsymbol{\sigma}_t^2) &= \mathbf{A}'^T \mathbf{h}_{t-1} + \mathbf{B}'^T \mathbf{v}_{t-1} + \mathbf{e}'.
 \end{aligned}$$

Multiple single-layer TSBNs can be stacked to

form a deep architecture. For a multi-layer TSBN, the hidden states of the lower layers are conditioned on the states of the higher layers. The connection structure of a multi-layer TSBN is given in Figure 2(e). The author investigated both the stochastic and deterministic conditioning for multi-layer models and showed that the deep model with stochastic conditioning works better. Experiment results on the human motion prediction task show a multi-layer TSBN with stochastic conditioning outperforms variants of SRTRBM and RTRBM.

4.3 FNN and RNN based Models

One shortcoming of DTGMs for STSF is that the learning and inference processes are often complicated and time-consuming. Recently, there is a growing trend of using FNN and RNN based methods for STSF. Compared with DTGM, FNN and RNN based methods are simpler in the learning and inference processes. Moreover, these methods have achieved good performance in many STSF problems like video prediction [31], [80], [81], precipitation nowcasting [5], [82], traffic speed forecasting [17], [32], [33], human trajectory prediction [4], and human motion prediction [14]. In the following, we will review these methods.

4.3.1 Encoder-Forecaster Structure

Recall that the goal of STSF is to predict the length- L sequence in the future given the past observations. The *Encoder-Forecaster* (EF) structure [5], [7] first encodes the observations into a finite-dimensional vector or a probability distribution using the encoder and then generates the one-step-ahead prediction or multi-step predictions using the forecaster:

$$\begin{aligned} \mathbf{s} &= f(\mathcal{F}_t; \theta_1), \\ \hat{\mathbf{X}}_{t+1:t+L} &= g(\mathbf{s}; \theta_2). \end{aligned} \quad (37)$$

Here, f is the encoder and g is the forecaster. In our definition, we also allow probabilistic encoder and forecaster, where \mathbf{s} and $\hat{\mathbf{X}}_{t+1:t+L}$ turn into random variables:

$$\begin{aligned} \mathbf{s} &\sim f(\mathcal{F}_t; \theta_1), \\ \hat{\mathbf{X}}_{t+1:t+L} &\sim \pi_g(\mathbf{s}; \theta_2). \end{aligned} \quad (38)$$

Using probabilistic encoder and forecaster is a way to handle uncertainty and we will review the details in Section 4.3.5. All of the surveyed FNN and RNN based methods fit into the EF framework. In the following, we will review how previous works design the encoder and the forecaster for different types of STSF problems, i.e., STSF-RG, STSF-IG, and TF-MPC.

4.3.2 Models for STSF-RG

For STSF-RG problems, the spatiotemporal sequences lie in a regular grid and can naturally be modeled by CNNs, which have proved effective for extracting

features from images and videos [62], [72]. Therefore, researchers have used CNNs to build the encoder and the forecaster. In [83], two input frames are encoded independently to two state vectors by a 2D CNN. The future frame is generated by linearly extrapolating these two state vectors and then mapping the extrapolated vector back to the image space with a series of 2D deconvolution and unpooling layers. Also, to better capture the spatial features, the author proposed the phase pooling layer, which keeps both the pooled values and pooled indices in the feature maps. In [6], multi-scale 2D CNNs with ReLU activations were used as the encoder and the forecaster. Rather than encoding the frames independently, the author concatenated the input frames along the channel dimension to generate a single input tensor. The input tensor is rescaled to multiple resolutions and encoded by a multi-scale 2D CNN. To better capture the spatiotemporal correlations, Vondrick et al. [9] proposed to use 3D CNNs as the encoder and forecaster. Also, the author proposed to use a 2D CNN to predict the background image and only use the 3D CNN to generate the foreground. Later, Shi et al. [82] compared the performance of 2D CNN and 3D CNN for precipitation nowcasting and showed that 3D CNN outperformed 2D CNN in the experiments. Kalchbrenner et al. [84] proposed the *Video Pixel Network* (VPN) that uses multiple causal convolution layers [85] in the forecaster. Unlike the previous 2D CNN and 3D CNN models, of which the forecaster directly generates the multi-step predictions, VPN generates the elements in the spatiotemporal sequence one-by-one with the help of causal convolution. Experiments show that VPN outperforms the baseline which uses a 2D CNN to generate the predictions directly. Recently, Xu et al. [86] proposed the PredCNN network that stacks multiple dilated causal convolution layers to encode the input frames and predict the future. Experiments show that PredCNN achieves state-of-the-art performance in video prediction.

Besides using CNNs, researchers have also used RNNs to build models for STSF-RG. The advantage of using RNN for STSF problems is that it naturally models the temporal correlations within the data. Srivastava et al. [7] proposed to use multi-layer LSTM networks as the encoder and the forecaster. To use the LSTM, the author directly flattened the images into vectors. Oh et al. [24] proposed to use 2D CNNs to encode the image frames before feeding into LSTM layers. However, the state-state connection structure of *Fully-Connected LSTM* (FC-LSTM) is redundant for STSF and is not suitable for capturing the spatiotemporal correlations in the data [5]. To solve the problem, Shi et al. [5] proposed the *Convolutional LSTM* (ConvLSTM). ConvLSTM uses convolution instead of full-connection in both the feed-forward and recurrent connections of LSTM. The formula of ConvLSTM is given in the

following:

$$\begin{aligned}
\mathcal{I}_t &= \sigma(\mathcal{W}_{xi} * \mathcal{X}_t + \mathcal{W}_{hi} * \mathcal{H}_{t-1} + b_i), \\
\mathcal{F}_t &= \sigma(\mathcal{W}_{xf} * \mathcal{X}_t + \mathcal{W}_{hf} * \mathcal{H}_{t-1} + b_f), \\
\mathcal{C}_t &= \mathcal{F}_t \circ \mathcal{C}_{t-1} + \\
&\quad \mathcal{I}_t \circ \tanh(\mathcal{W}_{xc} * \mathcal{X}_t + \mathcal{W}_{hc} * \mathcal{H}_{t-1} + b_c), \\
\mathcal{O}_t &= \sigma(\mathcal{W}_{xo} * \mathcal{X}_t + \mathcal{W}_{ho} * \mathcal{H}_{t-1} + b_o), \\
\mathcal{H}_t &= \mathcal{O}_t \circ \tanh(\mathcal{C}_t).
\end{aligned} \tag{39}$$

Here, all the states, cells, and gates are three-dimensional tensors. The author used two ConvLSTM layers as the encoder and the forecaster and showed that 1) ConvLSTM outperforms FC-LSTM when applied to the precipitation nowcasting problem and 2) the two-layer model performs better than the single-layer model. Similar to the way of extending LSTM to ConvLSTM, other types of RNNs, like GRU, can be extended to *Convolutional RNNs* (ConvRNNs) by using convolution in state-state transitions. Since ConvRNN combines the advantage of CNN and RNN, it is widely adopted in later researches [31], [80], [87], [88], [89] as a basic block for building DL models for STSF-RG. There are also attempts to improve the structure of ConvRNN. Shi et al. [82] proposed the *Trajectory GRU* (TrajGRU) model that improves ConvGRU by actively learning the recurrent connection structure. Because the convolutional recurrence structure adopted in ConvRNNs is location-invariant, it is not suitable for modeling location-variant motions, e.g., rotation and scaling. Thus, TrajGRU uses a learned subnetwork to output the recurrent connection structures:

$$\begin{aligned}
\mathcal{U}_t, \mathcal{V}_t &= \gamma(\mathcal{X}_t, \mathcal{H}_{t-1}), \\
\mathcal{Z}_t &= \sigma(\mathcal{W}_{xz} * \mathcal{X}_t + \sum_{l=1}^L \mathcal{W}_{hz}^l * \mathcal{H}_{t-1, \mathcal{U}_{t,l}, \mathcal{V}_{t,l}}), \\
\mathcal{R}_t &= \sigma(\mathcal{W}_{xr} * \mathcal{X}_t + \sum_{l=1}^L \mathcal{W}_{hr}^l * \mathcal{H}_{t-1, \mathcal{U}_{t,l}, \mathcal{V}_{t,l}}), \\
\mathcal{H}'_t &= f(\mathcal{W}_{xh} * \mathcal{X}_t + \\
&\quad \mathcal{R}_t \circ (\sum_{l=1}^L \mathcal{W}_{hh}^l * \mathcal{H}_{t-1, \mathcal{U}_{t,l}, \mathcal{V}_{t,l}})), \\
\mathcal{H}_t &= (1 - \mathcal{Z}_t) \circ \mathcal{H}'_t + \mathcal{Z}_t \circ \mathcal{H}_{t-1}.
\end{aligned} \tag{40}$$

Here, γ is a convolutional subnetwork that outputs L couples of flow fields $\mathcal{U}_t, \mathcal{V}_t \in \mathbb{R}^{L \times H \times W}$. $\mathcal{H}_{t-1, \mathcal{U}_{t,l}, \mathcal{V}_{t,l}}$ means to shift the elements in \mathcal{H}_{t-1} through the flow field $\mathcal{U}_{t,l}, \mathcal{V}_{t,l}$ by bilinear warping [90]. The connection structures of ConvRNN and TrajRNN are illustrated in Figure 3. Experiments on both the synthetic dataset and the real-world HKO-7 benchmark show that TrajGRU outperforms 2D CNN, 3D CNN, and ConvGRU. Wang et al. [91] proposed the *Predictive RNN* (PredRNN), which extends ConvLSTM by adding a new type of spatiotemporal mem-

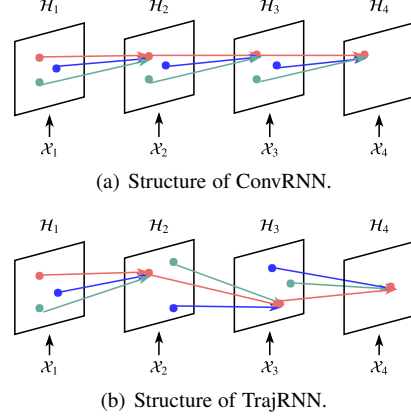


Fig. 3. Illustration of the connection structures of ConvRNN, TrajRNN. Links with the same color share the same transition weights. Source: [82].

ory. Unlike ConvLSTM, of which the memory states are constrained inside each LSTM layer, PredRNN maintains another global memory state to store the spatiotemporal information in the lower layers and in the previous timestamps. In PredRNN, memory states are allowed to zigzag across the RNN layers. The structure of PredRNN is illustrated in Figure 4(b). Here, \mathcal{M}_t^l s are the global spatiotemporal memory states and are updated in a similar way as the cell states in ConvLSTM except that it is updated in a zigzag order. Based on PredRNN, Wang et al. [81] proposed the PredRNN++ structure which added more nonlinearities when updating \mathcal{M}_t^l s. Experiments show that PredRNN++ outperforms both PredRNN and other baseline models including TrajGRU and variants of ConvLSTM.

Apart from studying the basic building blocks like CNNs and RNNs, another line of research tries to disentangle the motion and content of the spatiotemporal sequence to better forecast the future. Jia et al. [80] and Finn et al. [31] proposed to predict how the frames will transform instead of directly predicting the future frames. Finn et al. [31] proposed two transformation methods named *Convolutional Dynamic Neural Advection* (CDNA) and *Spatial Transformer Predictors* (STP). To predict the frame at timestamp $t + 1$, CDNA generates N 2D convolutional filters $\{\mathbf{W}_t^1, \mathbf{W}_t^2, \dots, \mathbf{W}_t^N\}$ and a masking tensor $\mathcal{M} \in \mathbb{R}^{N \times H \times W}$ where $\sum_{c=1}^N \mathcal{M}_{c,i,j} = 1$. The previous frame \mathcal{X}_t is transformed N times by these filters and these N transformed frames are linearly combined by the masking tensor to generate the final prediction. STP is similar to CDNA except that the \mathbf{W}_t^n s become the parameters of different affine transformations [90]. Similar to CDNA and STP, Jia et al. [80] proposed the *Dynamic Filter Network* (DFN) that directly outputs the parameters of a $K \times K$ local filter for every

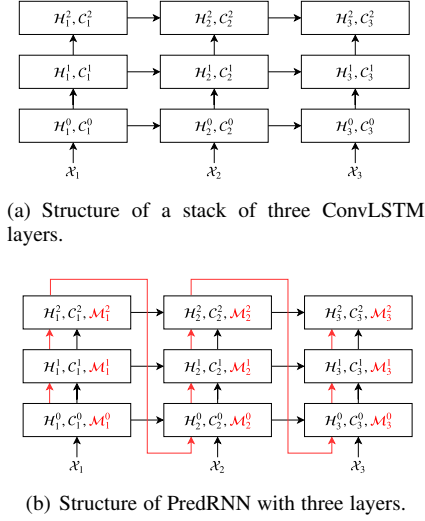


Fig. 4. Comparison of multi-layer ConvLSTMs and the PredRNN. \mathcal{M}_t is the global spatiotemporal memory that is updated along the red links.

location. The prediction is generated by applying these filters to the previous frame. Actually, DFN can be viewed as a special case of CDNA. If we manually fix the 2-dimensional convolutional filters in CDNA to have K^2 elements, of which only a single element is non-zero, CDNA turns into DFN. Villegas et al. [87] proposed *Motion-Content Network* (MCnet) which uses two separated networks to encode the motion and the content. In MCnet, the input of the motion encoder is the sequence of difference images $\{\mathcal{X}_2 - \mathcal{X}_1, \dots, \mathcal{X}_t - \mathcal{X}_{t-1}\}$ and the input of the content encoder is the last observed image \mathcal{X}_t . The encoded motion features and content features are fused together to generate the prediction. Experiments show that disentangling the motion and the content helps improve the performance. This architecture has been used in other papers [89], [92] to enhance the prediction performance.

4.3.3 Models for STSF-IG

Unlike STSF-RG problems, the measurements in STSF-IG problems are observed in some sparsely distributed stations which are more challenging to deal with. The simple strategy is to model the measurements observed at each station independently, which has been adopted in Yu et al. [17]. However, this strategy has not considered the correlation among different stations. A recent trend in solving this problem is to convert these sparsely distributed stations to a graph and utilize graph convolution to build the encoder and the forecaster. We will mainly introduce these graph convolution-based methods in this section.

Li et al. [32] proposed the *Diffusion Convolutional Recurrent Neural Network* (DCRNN) for traffic speed

forecasting. To convert the traffic stations into a graph, the author generated the adjacency matrix based on the thresholded pairwise road network distances. After that, DCRNN is used to encode the sequence of graphical data and predict the future. The formula of DCRNN is similar to ConvRNN except that the *Diffusion Convolution* (DC) is used in the input-state and state-state transitions. DC is a type of graph convolution that considers the direction information in the graph, which has the following formula:

$$\mathbf{Y} = \sum_{k=0}^{K-1} (\theta_{k,1}(\mathbf{D}_O^{-1}\mathbf{A})^k + \theta_{k,2}(\mathbf{D}_I^{-1}\mathbf{A}^T)^k)\mathbf{X}, \quad (41)$$

where $\mathbf{X} \in \mathbb{R}^{N \times C}$ is the input features, \mathbf{A} is the adjacency matrix, \mathbf{D}_O is the diagonal out-degree matrix, \mathbf{D}_I is the diagonal in-degree matrix, $\theta_{k,1}, \theta_{k,2}$ are the parameters, K is the number of diffusion steps, N is the number of nodes in the graph, and C is the number of the features. Unlike other graph convolution operators like ChebNet [69] which operates on undirected graphs, DC operates on a directed graph and is more suitable for real-life scenarios. Experiments on two real-world traffic datasets show that DCRNN outperforms FC-LSTM and the ChebNet-based *Graph Convolutional Recurrent Neural Network* (GCRNN) [32]. Later, Zhang et al. [33] proposed the *Graph GRU* (GGRU) network that generalizes the idea of DCRNN. Instead of proposing a single model, the author proposed a unified method for constructing an RNN based on an arbitrary graph convolution operator. The formula of GGRU is given in the following:

$$\begin{aligned} \mathbf{U}_t &= \sigma(\Gamma_{\Theta_{xu}}(\mathbf{X}_t, \mathbf{X}_t; \mathcal{G}) \\ &\quad + \Gamma_{\Theta_{hu}}(\mathbf{X}_t \oplus \mathbf{H}_{t-1}, \mathbf{H}_{t-1}; \mathcal{G})), \\ \mathbf{R}_t &= \sigma(\Gamma_{\Theta_{xr}}(\mathbf{X}_t, \mathbf{X}_t; \mathcal{G}) \\ &\quad + \Gamma_{\Theta_{hr}}(\mathbf{X}_t \oplus \mathbf{H}_{t-1}, \mathbf{H}_{t-1}; \mathcal{G})), \\ \mathbf{H}'_t &= h(\Gamma_{\Theta_{xh}}(\mathbf{X}_t, \mathbf{X}_t; \mathcal{G}) \\ &\quad + \mathbf{R}_t \circ \Gamma_{\Theta_{hh}}(\mathbf{X}_t \oplus \mathbf{H}_{t-1}, \mathbf{H}_{t-1}; \mathcal{G})), \\ \mathbf{H}_t &= (1 - \mathbf{U}_t) \circ \mathbf{H}'_t + \mathbf{U}_t \circ \mathbf{H}_{t-1}. \end{aligned} \quad (42)$$

Here, $\mathbf{U}_t, \mathbf{R}_t$ are the update gate and reset gate that controls the memory flow, \mathbf{H}_t is the hidden state, \mathbf{X}_t is the input, $h(\cdot)$ is the activation function, and $\Gamma(\cdot, \cdot)$ is an arbitrary graph convolution operator. The author also proposed a new multi-head gated graph attention aggregator that uses neural attention to aggregate the neighboring features. Experiments show that GGRU outperforms DCRNN if the gated attention aggregator is used. Yu et al. [93] proposed the *Spatio-Temporal Graph Convolutional Network* (STGCN) that applies graph convolution in a different manner. The author directly stacked multiple *Spatio-Temporal Convolution* (ST-Conv) blocks to build the network. The ST-Conv block is a concatenation of two temporal convolution

layers and one graph convolution layer. For the graph convolution operator, the paper tested the ChebNet and its first-order approximation [68] and found that ChebNet performed better. The major advantage of STGCN over RNN-based methods is it is faster in the training phase.

4.3.4 Models for TF-MPC

The goal of TF-MPC is to predict the future trajectories of all objects in the moving point cloud. Fragkiadaki et al. [13] proposed a basic RNN based method for this problem. In the model, the coordinate sequence of each human joint is encoded by the same LSTM. The drawback of this basic method is that it does not model the correlation among the individuals. To solve the problem, Alahi et al. proposed the SocialLSTM model [4] that uses a set of interacted LSTMs to model the trajectory of the crowd. In SocialLSTM, the position of each person is modeled by an LSTM and a social pooling layer interconnects these LSTMs. Social pooling aggregates the hidden states of the nearby LSTMs at the previous timestamp and use this aggregated state to help decide the current state. The transition rule of the SocialLSTM is given in the following:

$$\begin{aligned}
 \mathcal{N}_t^i &= \{j \mid j \in \mathcal{N}_i, |x_t^j - x_t^i| \leq m, |y_t^j - y_t^i| \leq n\}, \\
 \mathbf{H}_t^i &= \sum_{j \in \mathcal{N}_t^i} \mathbf{h}_{t-1}^j, \\
 \mathbf{e}_t^i &= \phi_1(x_t^i, y_t^i; \mathbf{W}_e), \\
 \mathbf{a}_t^i &= \phi_2(\mathbf{H}_t^i; \mathbf{W}_l), \\
 \mathbf{h}_t^i, \mathbf{c}_t^i &= \text{LSTM}(\mathbf{h}_{t-1}^i, \mathbf{c}_{t-1}^i, \text{vec}(\mathbf{e}_t^i, \mathbf{a}_t^i); \mathbf{W}_l).
 \end{aligned} \tag{43}$$

Here, \mathbf{H}_t^i is the aggregated state vector in the social pooling step, (x_t^j, y_t^j) is the coordinate of the j th person at timestamp t , ϕ_1, ϕ_2 are mapping functions, and m, n are distance thresholds. Experiments show that adding the social pooling layer is helpful for the TF-MPC task. In [94], the author proposed an improved version of social pooling called *View Frustum Of Attention* (VFOA) social pooling. This method defines the neighborhood set in the social pooling layer based on the calculated VFOA interest region. Jain et al. [14] proposed the *Structural-RNN* (S-RNN) network to solve the human motion prediction problem. S-RNN imposes structures in the recurrent connections based on a given spatiotemporal graph. For human motion prediction, this spatiotemporal graph is constructed by the relationship between different human joints. S-RNN maintains two types of RNNs, nodeRNN and edgeRNN. The outputs of edgeRNNs are fed to the corresponding nodeRNNs. Experiments show that the S-RNN achieves the state-of-the-art result in the human motion prediction problem.

4.3.5 Methods for Handling Uncertainty

Most real-world dynamical systems, e.g., atmosphere, are inherently stochastic and unpredictable. Simply assuming the outcome is deterministic, which is common in FNN and RNN based models, will generate blurry predictions. Thus, recent studies [9], [83], [95] have proposed ways for handling uncertainty.

In [13], the author proposed to use a probabilistic forecaster that outputs the parameters of a *Gaussian Mixture Model* (GMM). This enables the network to generate stochastic predictions. The author found that the smallest forecasting error was always produced by the most probable sample, which was similar to the output of a model trained by the Euclidean loss. Apart from using a probabilistic forecaster, the author also proposed a curriculum learning strategy which gradually adds noise to the input during training. This can be viewed as regularizing the model not to be overconfident about its predictions. This technique has also been adopted in later works [14]. Goroshin et al. [83] proposed to add a latent variable δ in the network to represent the uncertainty of the prediction. The latent variable δ_i is not determined by the input \mathcal{X}_i and can be adjusted in the learning process to minimize the loss. Before updating the parameters of the network, the algorithm first updates δ_i for k steps where k is a hyperparameter. At test time, the δ_i s are sampled based on its distribution on the training set.

Another approach to handle uncertainty is to utilize the recently popularized conditional *Generative Adversarial Network* (GAN) [96], [97]. Mathieu et al. [6] proposed to use conditional GAN as the loss function to train the multi-scale 2D CNN. The GAN loss used in the paper is given as follows:

$$\begin{aligned}
 \min_{\theta_G} \max_{\theta_D} \mathbb{E}_{\mathcal{Y} \sim p_{\text{data}}(\mathcal{Y}|\mathcal{X})} [\log D_{\theta_D}(\mathcal{X}, \mathcal{Y})] \\
 + \mathbb{E}_{z \sim p_z} [\log(1 - D_{\theta_D}(\mathcal{X}, G_{\theta_G}(\mathcal{X}, z)))].
 \end{aligned} \tag{44}$$

Here, D_{θ_D} is the discriminator, G_{θ_G} is the forecasting model, \mathcal{X} is the observed sequence, \mathcal{Y} is the ground-truth sequence, and $p(z)$ is the noise distribution. The discriminator is trained to differentiate between the ground-truth output and the prediction generated by the forecaster. On the other hand, the forecasting model is trained to fool the discriminator. The final loss function in the paper combined the GAN loss, the L2 loss, and the image gradient difference loss. Experiments show that the model trained with GAN generates sharper predictions. Vondrick et al. [9] used a similar GAN loss to train the 3D CNN. Jang et al. [89] proposed to use two discriminators to respectively discriminate the motion and appearance part of the generated samples.

Besides GAN, the *Variational Auto-encoder* (VAE) [98] has also been adopted for dealing with the uncertainty in STSF. Babaeizadeh et al. [95] proposed *Stochastic Variational Video Prediction* (SV2P) which

uses the following objective function:

$$l = -\mathbb{E}_{q_\phi(\mathbf{Z}|\mathbf{X}_{1:T})}[\log p_\theta(\mathbf{X}_{t:T} | \mathbf{X}_{1:t}, \mathbf{Z})] + D_{\text{KL}}(q_\phi(\mathbf{Z} | \mathbf{X}_{1:T}) || p(\mathbf{Z})), \quad (45)$$

where q_ϕ is the variational inference network that approximates the posterior of the latent variable \mathbf{Z} and p_θ is an RNN generator. The experiments show that adding VAE greatly alleviates the blurry problem of the predictions. Later, Denton & Fergus [99] proposed the *Stochastic Video Generator with Learned Prior* (SVG-LP) which uses a learned time-varying prior $p_\psi(\mathbf{Z}_t | \mathbf{X}_{1:t-1})$ to replace the fixed prior $p(\mathbf{Z})$ in (45). The extension in their paper is similar to the *Variational RNN* (VRNN) [100]. Learning a time-variant prior can be interpreted as learning a predictive model of uncertainty. The author showed in the experiments that using a learned prior improves the forecasting performance.

4.4 Remarks

In this section, we reviewed two types of DL based methods for STSF: DTGMs and FNN and RNN based methods. The majority of the reviewed DTGMs treat STSF as a general multivariate sequence forecasting problem and SRTRBM is the only model that gives special treatment of the spatiotemporal sequences. Also, learning a DTGM requires approximation and sampling techniques and is more complicated compared with the FNN and RNN based methods. However, DTGM can well capture the uncertainty in the data due to its stochastic generative process. On the other hand, the FNN and RNN based models are easy to train and fast for prediction. We reviewed the representative methods for STSF-RG, STSF-IG, and TF-MPC. Although these three types of STSF problems involve different tasks, the proposed methods for these problems have strong relationships. For example, ConvLSTM, DCRNN, and SocialLSTM, which are respectively proposed for STSF-RG, STSF-IG, and TF-MPC, improve upon RNN by introducing structured recurrent connections. To be more specific, ConvLSTM uses convolution in recurrent transitions, DCRNN uses the graph convolution, and SocialLSTM uses the social pooling layer. Therefore, the success of these models greatly relies on the appropriate design of the basic network architecture. Also, FNN and RNN based methods are usually deterministic and are not well-designed for capturing the uncertainty in the data. We thus reviewed some recently popularized techniques for handling uncertainty including GAN and VAE, which have improved the forecasting performance.

5 SUMMARY AND FUTURE WORKS

In this survey, we reviewed the machine learning based methods for STSF. We first examined the general strate-

gies for multi-step sequence forecasting, which contain the IMS, DMS, boosting strategy, and scheduled sampling. We showed that IMS and DMS have their advantages and disadvantages. The boosting strategy and scheduled sampling can be viewed as the mid-grounds between these two approaches. Next, we reviewed the specific methods for STSF. We divided these methods into two major categories: classical methods and DL based methods. As mentioned in 3.4 and 4.4, the existing classical methods and deep learning methods have their own trade-offs when solving the STSF problem. The overall summarization of the reviewed methods is given in Table 3.

There are many future research directions for machine learning for STSF. **1)** We can enhance the

strategies for the multi-step sequence forecasting. One direction is to design new training curriculum to shift from IMS to DMS. For example, we can extend the forecasting horizon exponentially from 1 to N to better model longer sequences, which is common in TF-MPC problems. Also, we can generalize the boosting method described in Section 2.4 to the general STSF problems. In addition, we can reduce the complexity of DMS from $O(h)$ to $O(\log h)$ by training $O(\log h)$ models that generate the predictions of horizons $2^0, 2^1, \dots, 2^{\log h - 1}$. We can use these models to make prediction for any horizon up to h in the spirit of binary coding. Moreover, the objective function mentioned in Section 2.5 will not be differentiable due to the sampling process and the original paper has just ignored this problem. This problem can be solved by approximating the gradient through techniques in reinforcement learning [101]. **2)** We can propose new

FNN and RNN based methods for STSF-IG problems. Compared with STSF-RG, DL for STSF-IG is less mature and we can bring the ideas in STSF-RG, like decomposing motion and content, adding a global memory structure, and using GAN/VAE to handle uncertainty, to STSF-IG. **3)** We can combine DTGMs

with FNN and RNN based methods to get the best of both worlds. The generation flavor of DTGMs has its role to play in FNN and RNN based models particularly for handling uncertainty. One feasible approach is to use the idea of *Bayesian Deep Learning* (BDL) [102], in which FNN and RNN based models are used for perception and DTGMs are used for inference. **4)** We

can study the type of TF-MPC problem where both the measurements and coordinates are changing. We can consider to solve the “video object prediction” problem where we have the bounding boxes of the objects at the previous frames and the task is to predict both the objects’ future appearances and bounding boxes. **5)** We

can study new learning scenarios like online learning and transfer learning for STSF. For the STSF problem, the online learning scenario is essential because spatiotemporal data, like regional rainfall, usually arrive in

TABLE 3

Summary of the reviewed methods for STSF. We put “√” under STSF-RG, STSF-IG, and TF-MPC to show the method has been extended to solve the problem. We put “√” under “Unc.” to show some of the methods have handled uncertainty.

Category	Subcategory	Methods	STSF-RG	STSF-IG	TF-MPC	Unc.
Classical	Feature-based	Spatiotemporal indicator [34]		√		
		Multiple predictors [3]		√		
	SSM	Group-based [40], [41]			√	√
		STARIMA [42], [43], [44]	√	√		√
		Low-rankness [46], [48], [49]	√	√		√
	GP	GP [16], [52], [53], [54]	√	√	√	√
DL	DTGM	TRBM [76]	√		√	√
		RTRBM [77]	√		√	√
		SRTRBM [78]	√	√	√	√
		TSBN [79]	√		√	√
		FC-LSTM	√	√	√	√
	FNN & RNN	2D CNN [6], [24] & 3D CNN [9]	√			√
		ConvLSTM [5] & TrajGRU [82]	√			
		PredRNN [81], [91]	√			
		VPN [84] & PredCNN [86]	√			
		Learn transformation [31], [80]	√			
		Motion + Content [87], [89], [92]	√			√
		DCRNN [32] & GGRU [33]		√		
		STGCN [93]		√		
		Social pooling [4], [94]			√	√
		S-RNN [14]			√	√
		δ -gradient [83]	√			√
		SV2P [95] & SVG-LP [99]	√			√

an online fashion [82]. The transfer learning scenario is also essential because STSF can act as a pretraining method. The model trained by STSF can be transferred to solve other tasks like classification and control [8], [31]. Designing a better transfer learning method for STSF will improve the performance of these tasks.

REFERENCES

- [1] B. R. Hunt, E. J. Kostelich, and I. Szunyogh, “Efficient data assimilation for spatiotemporal chaos: A local ensemble transform kalman filter,” *Physica D: Nonlinear Phenomena*, vol. 230, no. 1, pp. 112–126, 2007.
- [2] M. Ranzato, A. Szlam, J. Bruna, M. Mathieu, R. Collobert, and S. Chopra, “Video (language) modeling: A baseline for generative models of natural videos,” *CoRR*, vol. abs/1412.6604, 2014.
- [3] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li, “Forecasting fine-grained air quality based on big data,” in *KDD*, 2015.
- [4] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, “Social LSTM: Human trajectory prediction in crowded spaces,” in *CVPR*, 2016.
- [5] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, “Convolutional LSTM network: A machine learning approach for precipitation nowcasting,” in *NIPS*, 2015.
- [6] M. Mathieu, C. Couprie, and Y. LeCun, “Deep multi-scale video prediction beyond mean square error,” in *ICLR*, 2016.
- [7] N. Srivastava, E. Mansimov, and R. Salakhutdinov, “Unsupervised Learning of Video Representations using LSTMs,” in *ICML*, 2015.
- [8] C. Finn and S. Levine, “Deep visual foresight for planning robot motion,” in *ICRA*, 2017.
- [9] C. Vondrick, H. Pirsiavash, and A. Torralba, “Generating videos with scene dynamics,” in *NIPS*, 2016.
- [10] J. Walker, A. Gupta, and M. Hebert, “Dense optical flow prediction from a static image,” in *ICCV*, 2015.
- [11] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, “Who are you with and where are you going?” in *CVPR*, 2011.
- [12] G. W. Taylor, G. E. Hinton, and S. T. Roweis, “Modeling human motion using binary latent variables,” in *NIPS*, 2006.
- [13] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, “Recurrent network models for human dynamics,” in *ICCV*, 2015.
- [14] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena, “Structural-RNN: Deep learning on spatio-temporal graphs,” in *CVPR*, 2016.
- [15] A. Zonoozi, J. Kim, X. Li, and G. Cong, “Periodic-CRN: A convolutional recurrent model for crowd density prediction with recurring periodic patterns,” in *IJCAI*, 2018.
- [16] R. Senanayake, S. OCallaghan, and F. Ramos, “Predicting spatio-temporal propagation of seasonal influenza using variational gaussian process regression,” in *AAAI*, 2016.
- [17] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, “Deep learning: A generic approach for extreme condition traffic forecasting,” in *SDM*, 2017.
- [18] D. R. Cox, “Prediction by exponentially weighted moving averages and related methods,” *Journal of the Royal Sta-*

- tistical Society. Series B (Methodological)*, pp. 414–422, 1961.
- [19] G. Chevillon, “Direct multi-step estimation and forecasting,” *Journal of Economic Surveys*, vol. 21, no. 4, pp. 746–785, 2007.
 - [20] S. B. Taieb and R. Hyndman, “Boosting multi-step autoregressive forecasts,” in *ICML*, 2014.
 - [21] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, “Scheduled sampling for sequence prediction with recurrent neural networks,” in *NIPS*, 2015.
 - [22] P. Del Moral, A. Doucet, and A. Jasra, “Sequential monte carlo samplers,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 68, no. 3, pp. 411–436, 2006.
 - [23] J.-L. Lin and C. W. Granger, “Forecasting from non-linear models in practice,” *Journal of Forecasting*, vol. 13, no. 1, pp. 1–9, 1994.
 - [24] J. Oh, X. Guo, H. Lee, R. L. Lewis, and S. Singh, “Action-conditional video prediction using deep networks in atari games,” in *NIPS*, 2015.
 - [25] A. M. Lamb, A. G. ALIAS PARTH GOYAL, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio, “Professor forcing: A new algorithm for training recurrent networks,” in *NIPS*, 2016.
 - [26] I. Goodfellow, Y. Bengio, and A. Courville, “Deep learning,” 2016, book in preparation for MIT press. [Online]. Available: <http://www.deeplearningbook.org>
 - [27] M. Wytock and J. Z. Kolter, “Sparse gaussian conditional random fields: Algorithms, theory, and application to energy forecasting,” in *ICML*, 2013.
 - [28] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.
 - [29] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, “Curriculum learning,” in *ICML*, 2009.
 - [30] F. Huszar, “How (not) to train your generative model: Scheduled sampling, likelihood, adversary?” *CoRR*, vol. abs/1511.05101, 2015.
 - [31] C. Finn, I. Goodfellow, and S. Levine, “Unsupervised learning for physical interaction through video prediction,” in *NIPS*, 2016.
 - [32] Y. Li, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” in *ICLR*, 2018.
 - [33] J. Zhang, X. Shi, J. Xie, H. Ma, I. King, and D.-Y. Yeung, “GaAN: Gated attention networks for learning on large and spatiotemporal graphs,” in *UAI*, 2018.
 - [34] O. Ohashi and L. Torgo, “Wind speed forecasting using spatio-temporal indicators,” in *ECAI*, 2012.
 - [35] K. P. Murphy, *Machine learning: A probabilistic perspective*. MIT press, 2012.
 - [36] J. G. De Gooijer and R. J. Hyndman, “25 years of time series forecasting,” *International Journal of Forecasting*, vol. 22, no. 3, pp. 443–473, 2006.
 - [37] M. Fraccaro, S. r. K. Sønderby, U. Paquet, and O. Winther, “Sequential neural models with stochastic layers,” in *NIPS*, 2016.
 - [38] K. H. Brodersen, F. Gallusser, J. Koehler, N. Remy, S. L. Scott *et al.*, “Inferring causal impact using bayesian structural time-series models,” *The Annals of Applied Statistics*, vol. 9, no. 1, pp. 247–274, 2015.
 - [39] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.
 - [40] A. Asahara, K. Maruyama, A. Sato, and K. Seto, “Pedestrian-movement prediction based on mixed markov-chain model,” in *SIGSPATIAL*, 2011.
 - [41] W. Mathew, R. Raposo, and B. Martins, “Predicting future locations with hidden markov models,” in *UbiComp*. ACM, 2012.
 - [42] A. Cliff and J. Ord, “Model building and the analysis of spatial pattern in human geography,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 37, no. 3, pp. 297–348, 1975.
 - [43] A. Cliff and J. K. Ord, “Space-time modelling with an application to regional forecasting,” *Transactions of the Institute of British Geographers*, pp. 119–128, 1975.
 - [44] P. E. Pfeifer and S. J. Deutsch, “A starima model-building procedure with application to description and regional forecasting,” *Transactions of the Institute of British Geographers*, pp. 330–349, 1980.
 - [45] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: Forecasting and control*. John Wiley & Sons, 2015.
 - [46] M. T. Bahadori, Q. R. Yu, and Y. Liu, “Fast multivariate spatio-temporal analysis via low rank tensor learning,” in *NIPS*, 2014.
 - [47] B. Recht, M. Fazel, and P. A. Parrilo, “Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization,” *SIAM Review*, vol. 52, no. 3, pp. 471–501, 2010.
 - [48] R. Yu, D. Cheng, and Y. Liu, “Accelerated online low-rank tensor learning for multivariate spatio-temporal streams,” in *ICML*, 2015.
 - [49] R. Yu and Y. Liu, “Learning from multiway data: Simple and efficient tensor regression,” in *ICML*, 2016.
 - [50] A. G. Journel and C. J. Huijbregts, *Mining geostatistics*. Academic press, 1978.
 - [51] C. E. Rasmussen, “Gaussian processes for machine learning,” 2006.
 - [52] P. Trautman and A. Krause, “Unfreezing the robot: Navigation in dense, interacting crowds,” in *IROS*, 2010.
 - [53] S. Flaxman, “Machine learning in space and time,” PhD dissertation, Carnegie Mellon University, 2015.
 - [54] S. Flaxman, A. G. Wilson, D. B. Neill, H. Nickisch, and A. J. Smola, “Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods,” in *ICML*, 2015.
 - [55] P. Smolensky, “Information processing in dynamical systems: Foundations of harmony theory,” DTIC Document, Tech. Rep., 1986.
 - [56] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Neural Computation*, vol. 14, no. 8, pp. 1771–1800, 2002.
 - [57] M. Welling, M. Rosen-Zvi, and G. E. Hinton, “Exponential family harmoniums with an application to information retrieval,” in *NIPS*, 2004.
 - [58] X. Glorot, A. Bordes, and Y. Bengio, “Deep sparse rectifier neural networks,” in *AISTATS*, 2011.
 - [59] A. L. Maas, A. Y. Hannun, and A. Y. Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *ICML*, 2013.
 - [60] A. Mnih and K. Gregor, “Neural variational inference and learning in belief networks,” in *ICML*, 2014.
 - [61] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, “The “wake-sleep” algorithm for unsupervised neural networks,” *Science*, vol. 268, no. 5214, p. 1158, 1995.
 - [62] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *ICCV*, 2015.
 - [63] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in *ICLR*, 2016.
 - [64] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” in *CVPR*, 2010.
 - [65] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *ECCV*, 2014.
 - [66] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, “Spectral networks and locally connected networks on graphs,” in *ICLR*, 2014.
 - [67] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams,

- “Convolutional networks on graphs for learning molecular fingerprints,” in *NIPS*, 2015.
- [68] T. N. Kipf and M. Welling, “Semi-supervised classification with graph convolutional networks,” in *ICLR*, 2017.
- [69] M. Defferrard, X. Bresson, and P. Vandergheynst, “Convolutional neural networks on graphs with fast localized spectral filtering,” in *NIPS*, 2016.
- [70] W. Hamilton, Z. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *NIPS*, 2017.
- [71] F. Rosenblatt, *Perceptions and the theory of brain mechanisms*. Spartan books, 1962.
- [72] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in *NIPS*, 2012.
- [73] J. Nocedal and S. Wright, *Numerical optimization*. Springer Science & Business Media, 2006.
- [74] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training recurrent neural networks,” in *ICML*, 2013.
- [75] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [76] I. Sutskever and G. E. Hinton, “Learning multilevel distributed representations for high-dimensional sequences,” in *AISTATS*, 2007.
- [77] I. Sutskever, G. E. Hinton, and G. W. Taylor, “The recurrent temporal restricted boltzmann machine,” in *NIPS*, 2009.
- [78] R. Mittelman, B. Kuipers, S. Savarese, and H. Lee, “Structured recurrent temporal restricted boltzmann machines,” in *ICML*, 2014.
- [79] Z. Gan, C. Li, R. Henao, D. E. Carlson, and L. Carin, “Deep temporal sigmoid belief networks for sequence modeling,” in *NIPS*, 2015.
- [80] X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, “Dynamic filter networks,” in *NIPS*, 2016.
- [81] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, “PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning,” in *ICML*, 2018.
- [82] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. WOO, “Deep learning for precipitation nowcasting: A benchmark and a new model,” in *NIPS*, 2017.
- [83] R. Goroshin, M. F. Mathieu, and Y. LeCun, “Learning to linearize under uncertainty,” in *NIPS*, 2015.
- [84] N. Kalchbrenner, A. v. d. Oord, K. Simonyan, I. Danihelka, O. Vinyals, A. Graves, and K. Kavukcuoglu, “Video pixel networks,” in *ICML*, 2016.
- [85] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, “WaveNet: A generative model for raw audio,” *CoRR*, vol. abs/1609.03499, 2016.
- [86] Z. Xu, Y. Wang, M. Long, and J. Wang, “PredCNN: Predictive learning with cascade convolutions,” in *IJCAI*, 2018.
- [87] R. Villegas, J. Yang, S. Hong, X. Lin, and H. Lee, “Decomposing motion and content for natural video sequence prediction,” in *ICLR*, 2017.
- [88] A. Bhattacharyya, B. Schiele, and M. Fritz, “Accurate and diverse sampling of sequences based on a best of many sample objective,” in *CVPR*, 2018.
- [89] Y. Jang, G. Kim, and Y. Song, “Video prediction with appearance and motion conditions,” in *ICML*, 2018.
- [90] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, “Spatial transformer networks,” in *NIPS*, 2015.
- [91] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, “PredRNN: Recurrent neural networks for predictive learning using spatiotemporal lstrms,” in *NIPS*, 2017.
- [92] J. Xu, B. Ni, Z. Li, S. Cheng, and X. Yang, “Structure preserving video prediction,” in *CVPR*, 2018.
- [93] B. Yu, H. Yin, and Z. Zhu, “Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting,” in *IJCAI*, 2018.
- [94] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani, “MX-LSTM: Mixing tracklets and vislets to jointly forecast trajectories and head poses,” in *CVPR*, 2018.
- [95] M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, “Stochastic variational video prediction,” in *ICLR*, 2018.
- [96] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *NIPS*, 2014.
- [97] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” *CoRR*, vol. abs/1411.1784, 2014.
- [98] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *ICLR*, 2014.
- [99] E. Denton and R. Fergus, “Stochastic video generation with a learned prior,” in *ICML*, 2018.
- [100] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, “A recurrent latent variable model for sequential data,” in *NIPS*, 2015.
- [101] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour *et al.*, “Policy gradient methods for reinforcement learning with function approximation,” in *NIPS*, 1999.
- [102] H. Wang and D.-Y. Yeung, “Towards bayesian deep learning: A framework and some existing methods,” *IEEE Transaction on Knowledge and Data Engineering*, vol. 28, no. 12, pp. 3395–3408, Dec. 2016.

PLACE
PHOTO
HERE

Fellowship in 2014.

PLACE
PHOTO
HERE

Xingjian Shi received his BEng degree in information security from Shanghai Jiao Tong University, China. He is currently a Ph.D student at the Department of Computer Science and Engineering of Hong Kong University of Science and Technology. His research interests are in deep learning and spatiotemporal analysis. He received the Hong Kong PhD

Dit-Yan Yeung received his BEng degree in electrical engineering and MPhil degree in computer science from the University of Hong Kong, and PhD degree in computer science from the University of Southern California. He started his academic career as an assistant professor at the Illinois Institute of Technology in Chicago. He then joined the Hong Kong University of Science and Technology where he is now a full professor and the Acting Head of the Department of Computer Science and Engineering. His research interests are in computational and statistical approaches to machine learning and artificial intelligence.