CAU: A Causality Attention Unit for Spatial-temporal Sequence Forecast

Bo Qin, Fanqing Meng, Shijin Yuan, and Bin Mu

***Abstract*—Existing convolution recurrent neural networks (ConvRNNs)-based memory cells majorly take advantage of gated structures and attention mechanisms to extract discontinuous latent associations for spatial-temporal sequence forecast (STSF) problems, which may lead to serious over-fitting and spurious relationships with correlated noise. It is a consensus that incorporating cause-effect relationships in modeling can alleviate these problems. In this paper, we propose a Causality Attention Unit (CAU) to assist ConvRNNs by complementing the causal inference ability in a plug-and-play way. Specifically, CAU serially consists of the attention module and causality module. The former is constructed by a spatial-channel attention layer, which preliminarily generates the correlated future with the correlations between historical memories and the current state. The latter borrows the idea of transfer entropy (*TE*) to detect the latent cause-effect relationships and precisely correct the correlated future. A space-time exchange strategy for accelerating the calculation of *TE* in CAU is also designed. CAU can be easily combined with the existing ConvRNN cells, and we construct a simple general model to predict long-term spatial-temporal series, which consists of encoder/decoder and stacked CAU paralleled to stacked ConvRNN cells. After determining the optimal model structure, we carry out a series of experiments to evaluate model performance, including comparisons with other advanced models, training loss analysis, and multiple ablation and sensitivity studies. Experimental results show that our proposed model can effectively improve the performances of existing ConvRNNs to the state-of-the-are level on representative public datasets, including Moving MNIST, KTH, BAIR, and WeatherBench. The ablation and sensitivity studies verify the superiority of CAU. The learned causal maps precisely distinguish the pixel attributions and motion characteristics in sophisticated entangled scenarios.**

***Index Terms*—Spatial-temporal Sequence Forecasting, Causality Attention Unit, Causal Inference, Transfer Entropy.**

I. Introduction

Spatial-temporal sequence forecasting (STSF) problem [1] is one of the most cutting-edge challenges, which manifests in multiple research areas, such as video prediction [2]–[4], traffic congestion estimation [5], motion/trajectory prediction [6], and even weather/climate forecasting [7]–[10], etc. In recent years, deluges of advanced and efficient deep learning (DL) frameworks are proposed emergently for tackling the STSF problem, which has achieved landmark progress and is driving our pursuit of more accurate forecasts.

The rapid developments of DL STSF models can be initially summarized as starting from the design of convolutional recurrent neural networks (ConvRNNs), e.g., convolution long short-term memory (ConvLSTM) [11] and convolution gated recurrent unit (ConvGRU) [12]. ConvRNN possesses not only the typical capability of extracting spatial characteristics via convolutions but retains the ability to infer the future via multiple gated structures (e.g., update, reset, and forecast gates) of RNN. Though such models are the successful expansions of conventional 1-dimensional LSTM to the 2-dimensional manifolds, they still inherit some shortcomings inevitably, such as the narrow temporal receptive field and restricted expressions for complex scenarios (See Section II.A).

Furthermore, with the development of attention mechanisms [13], researchers tend to leverage distinctive attentions to augment the capture of potential spatial dependencies and long-term temporal memories, e.g., Memory attention unit (MAU) [14], spatial-temporal attention based memory (STAM) [15], etc. These attentions are usually hand-craftily designed to focus on qualifying the detailed and crucial correlations between different spatial and temporal states, which raises the upper limits of describing the most significant variations in sequence evolutions and broadens the practicalities in the real world (See Section II.B).

The sophisticated variability of temporal memories and the elusive uncertainty of spatial distributions are indeed the toughest barriers to the STSF problem, which can be alleviated by customized attention mechanisms to some extent. However, excessive focus on the correlation dependencies in spatial-temporal contexts can lead to serious over-fitting of training data and capture spurious relationships with unpredictable noise [16], [17], especially when the objects in the scene are severely

This manuscript is submitted on Apr. 3, 2023; revised on Aug. 5, 2023; and accepted on Oct. 6, 2023.

This study is supported in part by the Meteorological Joint Funds of the National Natural Science Foundation of China (U2142211), in part by the National Key Research and Development Program of China (2020YFA0608000), in part by the National Natural Science Foundation of China (42075141), and in part by the the Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities. *(Corresponding author: Shijin Yuan and Bin Mu)*

Bo Qin is with the Department of Atmospheric and Oceanic Sciences & Institude of Atmospheric Sciences, Fudan University, Shanghai, 200438, P.R.China; also with the Key Laboratory of Polar Atmosphere-ocean-ice System for Weather and Climate, Ministry of Education. Fanqing Meng is with the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, 200240, P.R.China. Shijin Yuan and Bin Mu are with the School of Software Engineering, Tongji University, Shanghai, 201804, P.R.China; also with the State Key Laboratory of Intelligent Autonomous Systems. (e-mail: boqin@fudan.edu.cn; mengfanqing33@gmail.com; {yuanshijin; binmu}@tongji.edu.cn,).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org

entangled. For instance, the arms of a walking man are rotating and shifting, two pedestrians overlap each other in the view of a camera, and the passing car is moving away and shrinking. But the waving arms, as well as the overlapped pedestrians and the running car, are usually predicted very vaguely or even disappear by many attention-based models. This is because the correlations are usually non-directional, which cannot distinguish the causes and effects among disentangling abstract features (or representations) in the top-level (so-called encoder) of DL models[18], [19]. So, some significant information is mistaken for noise. Not to mention the captured associations may be useless (spurious) that cannot contribute to deducting the future.

To solve the above problems, numerous efforts have been paid on incorporating causality into the DL model. Causality has unambiguous directions and exists in extensive time-series data, especially non-stationary distributions, which is the key to further performance improvement. At present, there has been lots of pioneering work on causal mining using neural networks, but few studies have focused on the STSF forecasts so far (See Section II.C).

Therefore, unifying the causality in STSF modeling has great significance. Suppose there is a frame $\mathcal{X}_t$ that contains N grids $x_t^{(i)} \in \mathcal{X}_t$ $(i = 1{:}N)$ and evolves along the timeline t. For the prediction of a certain grid i, a direct formula $x_{t+1}^{(i)} = \mathcal{F}(x_{1:t}^{(i)})$ is the common approach ($\mathcal{F}(\cdot)$ is the forecasting system, and $1{:}t$ represents the historical sequence). But when the other grids $x_{1:t}^{(j)}$ $(j = 1{:}N\backslash i)$ also contribute to this prediction according to the dynamical derivation or physical analysis, it is better to append them as $x_{t+1}^{(i)} = \mathcal{F}(x_{1:t}^{(i)}, x_{1:t}^{(j)})$, which can effectively improve forecast accuracy especially when there is a cause-and-effect relationship between $x_{1:t}^{(j)}$ and $x_{t+1}^{(i)}$. Take a common scene in the Moving MNIST dataset as an example (See Fig. 1). The digit "5" and "7" are sliding and bouncing in a fixed region. When their previous locations are given in Fig. 1(a), the causal inference helps infer the future states more accurately. As visualized in Fig. 1(b) and (c), they are both monitoring the locations of each other when they are entangled, which is because they tend to distinguish the attribution of pixels during motions. In addition, they are both monitoring the boundary before they are about to bounce. These are the "causes" of their future variations, with which we can precisely predict the evolutions as shown in Fig. 1(d). (Note that the sub-figures in Fig. 1 are all the inference processes of our proposed model by visualizing the learned causal maps.)

The above illustrations also work well in other scenarios. Reasonable use of causality can identify and select the most valuable features from the historical state. In this paper, we propose a Causality Attention Unit (CAU) to mine latent cause-effect relationships overlooked by vanilla ConvRNNs according to the above notion. Specifically, there are two sequential modules in CAU: The attention module and the causality module. The attention module uses the historical memories and current state to preliminarily infer future

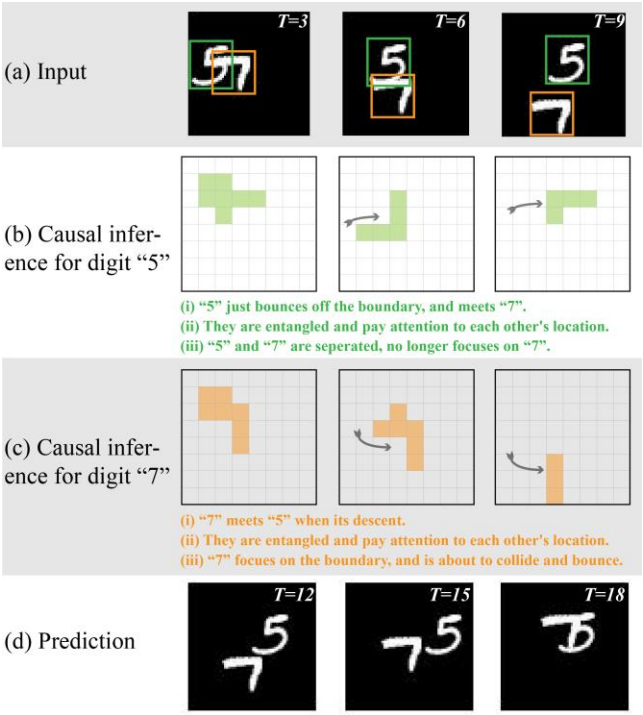

Fig. 1. An example of using causal inference in the STSF problem for Moving the MNIST dataset. (a) represents the historical status. (b) and (c) show the causal maps for the digit "5" and "7" respectively, which means they are monitoring the locations of each other for precisely distinguishing the pixel attribution and the location of the boundary before bouncing. The gray arrows describe the motion directions. (d) represents the prediction results. Note that all the sub-figures are the real results of our proposed model after refactoring and post-processing.

variations, and the causality module corrects such inferred variations with the help of transfer entropy (TE) mathematically, which is a concept from information theory and can be interpreted as the quantized cause-effect information (See Section III.C). Meanwhile, we also design a novel way of space-time exchange to accelerate the calculation of TE, which originally has multi-level loops. The computational efficiency drops from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$. CAU can be easily combined with the existing ConvRNN cells, and then we construct a simple general model, which consists of an encoder, a decoder, and stacked CAUs paralleled to stacked ConvRNN cells. This model can iteratively predict long-term spatial-temporal series. After determining the optimal model structure, we carry out a series of experiments on comparisons with other ConvRNN-based STSF models, and model performance evaluation via ablation and sensitivity studies.

Some scientific contributions are as follows:

- CAU can mine the latent causal relationships via TE mathematically, which can be rapidly calculated by a space-time exchange strategy and can be easily combined with the existing ConvRNN cell (e.g., MIM, PredRNN++) in a plug-and-play way.
- The simple general model based on CAU can effectively improve the performances of existing ConvRNNs compared to other advanced models on multiple public datasets, including common scenarios (Moving MNIST dataset), video predictions (KTH and BAIR datasets), and

weather forecasting (WeatherBench dataset).

- The simple general model based on CAU is easier to converge under the same training configurations because it has no gated structures, which weakens gradient explosion and vanishing.
- The visualized causal maps learned by CAU demonstrate that CAU can precisely distinguish the pixel attributions and motion characteristics in sophisticated entangled scenarios, such as rotating, shifting, and scaling in common STSF problems.

The remainder of the paper is organized as follows. Section II provides a summary of the literature regarding the existing STSF models. Section III depicts the methodology, including the formalization of the STSF problem, the framework of CAU, the interior implementation and accelerated calculation of high-dimensional TE, and a simple general model based on CAU. Section IV describes the experimental schemes, which consist of the introduction of datasets, the comparison results, and the performance evaluations. The conclusions and future works are summarized in Section V.

II. Related Works

In this section, we introduce the development footprints of DL STSF models from the most fundamental ConvRNN to attention-augmented frameworks. Subsequently, we also illustrate some outstanding methods of performing causal inference by neural networks.

A. ConvRNN-based STSF Models

Due to the irreplaceable capability of resolving temporal memories, RNN, e.g., LSTM and GRU, is the first structure naturally considered in the STSF problem. [11] designs the ConvLSTM by integrating the convolution and LSTM together, which has a very low time-/resource-consuming in dealing with spatial-temporal features. The same superiority is also reflected in ConvGRU [12]. Furthermore, to extend the effective temporal length (memories) that vanilla RNNs can handle, [20] proposes PredRNN by adding more memory units in LSTM cells, and [21] proposes PredRNN++ by adding a gradient highway unit (GHU), which are all conducive to the preservation and transmission of longer-term memories. In addition, other novel upgrades are also advantaged in dealing with non-stationary sequences, which incorporate spatial dependencies and incorporate physical laws. For example, [22] designs the Memory in Memory (MIM) module to individually resolve instantaneous and tendency. [23] proposes E3D-LSTM to capture spatial features in different timestamps by 3D convolution and increase the temporal receptive field by a new memory unit named eidetic. [24] introduces PhyDNet by constructing the simulations of dynamical partial differential equations paralleled to the memory unit, which effectively represents the prior physical knowledge.

Besides, there are many Transformer-based spatial-temporal forecasting models proposed recently [25]–[28], which have further advanced prediction accuracy and quality, demonstrating great potential. These models formalize the STSF as the relationships mining from historical sequence with no memory filtering/transferring involved during prediction iterations. In this paper, we majorly focus on the ConvRNN-based models. How to incorporating causality into Transformer-based models is the future topic.

B. Attention Techniques for Video-related Tasks

To overcome the common issues of information loss and gradient disappearance in vanilla ConvRNN-based models for multiple video-related tasks, (self-)attention techniques are gradually applied for capturing comprehensive correlations among different semantics (dimensions). Specifically, for STSF problem, SA-ConvLSTM (self-attention ConvLSTM) [29] is a successful attempt in quantifying the associations between the current state and historical memories, improving long-term and teleconnections. Some other upgrades [30] contain more efficient and diverse structures based on (self-)attention, which exhibit significant skills in the modulation of spatial-temporal relationships at the pixel level. The adjustable spatial and temporal receptive fields via attention mechanisms attract lots of interest. Memory attention unit (MAU) [14] and its variation, spatial-temporal attention based memory (STAM) [15], are two representative memory cells, which can both optimally extend the receptive fields. In addition, temporal attention unit (TAU) [31] decouples the attention into intra-frame statics and inter-frame dynamics, bringing a brand new insight into attention utilization and improving prediction performance. In addition, for other typical video-related tasks, such as action recognition and anomaly detection, multiple attention techniques are coupled in the temporal, pixel, or channel dimension to enhance the discrimination of locating the key frames of action occurrence and recognizing the key features of action type [32]–[34]. This shows that the customized attention mechanism, as a neural operator, can be applied to any dimension and exert positive effects in a plug-and-play way.

In this paper, the proposed CAU performs the spatial-temporal evolutions on the single-frame image alone the time lines. The internal feature maps in CAU do not contain temporal dimensions, so we choose the spatial-channel coupled attention mechanism, which can extract the multimodal (channel level) and pixel (spatial level) dependencies of the decoded feature map of a single-frame image simultaneously.

C. Causal Inference for Video-related Tasks

The most criticized points of (self-)attention are over-fitting and easy-capturing of spurious correlations [16], which can be alleviated by effective sparsity [35], [36] and causal inference [37]. The latter is a more direct way, which can make more concrete deductions that are not easy to change under sophisticated scenes. However, for the STSF problem, existing works are few and focus on integrating the (dilated) causal convolutions into STSF models [38], [39], which only consider the form but ignore the principle and measurement of cause-effect. Meanwhile, these models often encounter the shortcomings of local information loss, high computation complexity, and low interpretability. For the other video-related tasks, causal inference has been pioneeringly applied in model

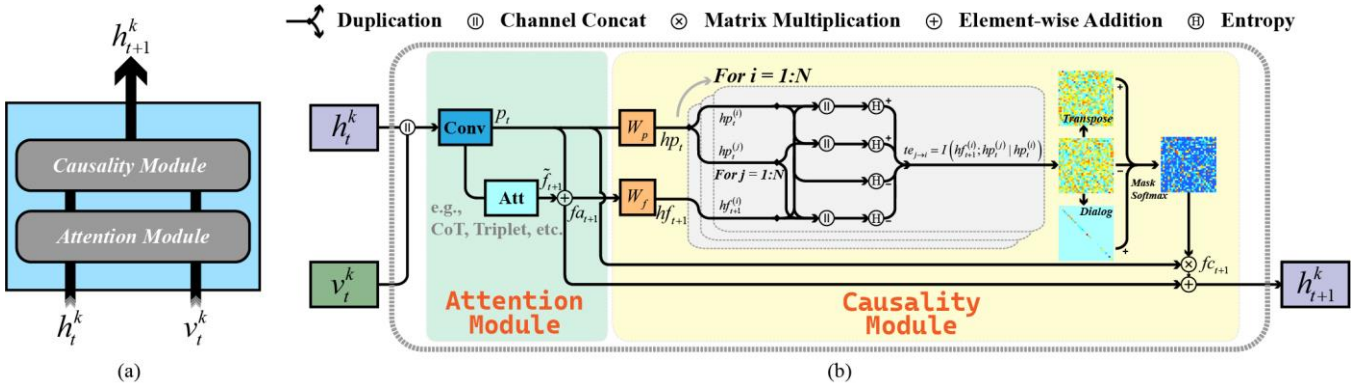

Fig. 2. (a) represents the two serial modules (attention module and causality module) of CAU. (b) shows the detailed calculation process of these two modules.

training strategy to effectively explain the decision process of "black-box" models on the given results. For example, [41] uses causal graphs to analyze the confounder effects of unsupervised training of pseudo-labels and eliminates the negative effects of errors/noises in pseudo-labels via blocking the backdoor effect paths, improving the performance of anomaly detection. [42] analyzes the spatial-temporal effects of the image-to-video adaption using a causal graph with counterfactual inference, and applies the learned spatial-temporal migration features to compensate for performance degradation during classifier migration. The causal effects involved in the above models generally act at the macro level, such as global temporal scale, appearance scale and action scale, and only implicitly express the causal relationships through the high degree of neural networks nonlinearity. Recent years, some causal inference approaches are proposed in the directed acyclic graph (DAG) learning in the areas of natural language processing (NLP), such as text generation [40], which is a big forward for sequence forecasting problems. However, due to the difficult unification of the textual token and image token, these advanced insights are hard to transfer into video-related tasks to perform causal inference at micro (image patch or pixel) level.

In this paper, the transfer entropy-based causal inference used in our proposed CAU quantifies the causality of individual pixel features of single-frame image within the spatial-temporal sequence, acting on a micro spatial scale with explicit mathematical implications.

III. Methodology

In this section, we first clarify the general formalization of the STSF problem. Then, we illustrate our proposed Causality Attention Unit (CAU), which is implemented via transfer entropy (TE) mathematically and can complement the existing methods with the ability of mining the latent causal relationships hidden in spatial-temporal sequence. Meanwhile, we design a space-time exchange strategy to rapidly calculate TE. Finally, we propose a simple general STSF model based on CAU and exhibit its detailed structure, especially the encoder and decoder.

A. STSF Problem Formalization

STSF problem can be typically regarded as using historical τ sequential data $\mathcal{X}_{1:t}$ to forecast the observed scene (ground truth) $\mathcal{X}_{t+1}$ of the next time step, which can be depicted in (1),

$$\widehat{\mathcal{X}}_{t+1} = \mathcal{F}(\mathcal{X}_{1:t}) \tag{1}$$

where t is the current time and $\widehat{\mathcal{X}}_{t+1}$ is the prediction result. $\mathcal{F}(\cdot)$ represents the forecasting system. In general, this paradigm can be iterated by multiple times by feeding the predicted results into the right side of this equation to obtain a predicted sequence $\widehat{\mathcal{X}}_{t+1:T}$. The goal of the STSF problem is to optimize (2)

$$\min \sum_{s=t+1}^{T} \mathcal{Q}\left(\mathcal{X}_s, \widehat{\mathcal{X}}_s\right) \tag{2}$$

where $\mathcal{Q}$ is the chosen quality assessment indicator, such as $SSIM$ (structural similarity) for human-eye perception and MSE (mean square error) for distance measure of errors.

DL model, as a common solution, is a reliable tool for constructing the forecasting system $\mathcal{F}$, which can be usually decoupled into three parts with different effects as (3).

$$\begin{cases} Encoder\colon v_t = \mathcal{E}(\mathcal{X}_t) \\ Informer\colon h_{t+1} = \mathcal{I}(h_t, v_t) \\ Decoder\colon \widehat{\mathcal{X}}_{t+1} = \mathcal{D}(h_{t+1}) \end{cases} \tag{3}$$

where $\mathcal{E}$ (Encoder) is used to capture the current spatial-temporal feature v_t, $\mathcal{I}$ (Informer) is used to make a memory inference for the hidden state h_{t+1} of the next sequential time step with hidden memory h_t and v_t (h_0 is generally obtained by internal initialization, and when the unit layer is more than 1, v_t is usually from the output of previous layer), and $\mathcal{D}$ (Decoder) is used to restore the feature to the predicted value $\widehat{\mathcal{X}}_{t+1}$. $\mathcal{E}$ and $\mathcal{D}$ are usually constructed by convolutional skeletons. ConvRNN, such as some advanced variants (e.g., PredRNN, MIM, etc.), plays the role of $\mathcal{I}$, which is a memory inference unit with varied gated structures as (4). $forget_t(\cdot,\cdot)$ serves as the forget gate and $input_t(\cdot,\cdot)$ represents the input gate ($*$ is convolution operator). These three parts modulate closely to achieve more accurate forecasts.

$$h_{t+1} = forget_t(v_t, h_t) * h_t + input_t(v_t, h_t) * v_t \tag{4}$$

$\mathcal{I}$ is the key component of the STSF problem. Broadly speaking, $forget_t(\cdot,\cdot)$ and $input_t(\cdot,\cdot)$ are both with activation functions in the interval [0,1], and are tied as $forget_t(\cdot,\cdot) + input_t(\cdot,\cdot) = 1$. However, such architecture makes the forget gate very easy to saturate (i.e., close to 1), especially when addressing long-term memories [43], which induces the gradient vanishing and hampers memory updating. Meanwhile, although multiple attention mechanisms are applied in $\mathcal{I}$ to enhance the extraction of discontinuous latent spatial-temporal dependencies, they tend to "create" fake associations between

two dependent variables especially when they are influenced by the third latent variable according to Reichenbach's common cause principle [19], which are spurious correlations. This is a very fatal issue for the STSF problem, because the spurious correlations may cause the feature of unrelated objects in the scene to be updated together after receiving the same high attention weights, making predictions blurry and distorted.

It is a consensus that complementing neural networks with the ability of causal inference can effectively eliminate useless attentions. The process of causal inference emphasizes the lagged spatial-temporal relationships and imperceptible causes, which is naturally suitable for predicting future scenarios in the STSF problem. So, in this paper, we propose a Causality Attention Unit (CAU) to achieve this purpose. It can perform the causal inference paralleled to ConvRNN in a plug-and-play way.

B. Causality Attention Unit: CAU

CAU has two sub-modules, which can be systematically described in (5) ($k = 1{:}\,K$ represents the unit layer).

$$\begin{cases} attention{:}\ fa_{t+1} = \mathcal{A}(h_t^k, v_t^k) \\ causality{:}\ fc_{t+1} = \mathcal{C}(h_t^k, v_t^k, fa_{t+1}) \\ output{:}\ h_{t+1}^k = fa_{t+1} + fc_{t+1} \end{cases} \tag{5}$$

where $\mathcal{A}$ and $\mathcal{C}$ are the attention and causality modules respectively. The former is to generate correlated future fa_{t+1} according to latent correlations, and the latter is to mine the cause-effect relationships hidden in historical sequence, generating causal future fc_{t+1}.

As (5), CAU has two internal components, which are connected sequentially. We model them respectively as shown in Fig. 2(a). Note that there is no gated structure in CAU, because CAU makes correlation and causality inferences along the timeline, which plays a substitute role in memory filtering and updating.

The detailed structures of attention (green shadow) and causality (yellow shadow) modules are described in Fig. 2(b), the operators of which are marked above the figure. Overall, before feeding inputs into the attention module, we first concatenate (marked as || in (7)) the hidden state h_t^k and current status v_t^k, which is a simple but effective manner to couple all memories. Then, after the propagation of these two modules, the summation output h_{t+1}^k is passed into the decoder $\mathcal{D}$ or the next layer $(k+1)$.

The attention module is also the summation of two parts as (6) (the green shadow of Fig. 2(b)),

$$fa_{t+1} = \mathcal{A}(h_t^k, v_t^k) = p_t + \tilde{f}_{t+1} \tag{6}$$

where p_t is the resolved previous (historical) memories and $\tilde{f}_{t+1}$ is the residual correlated future. These two parts can be obtained by (7) respectively, where $Att(\cdot)$ is the chosen attention layer, and $Conv(\cdot)$ is used to resolve the coupled memories.

$$p_t = Conv(h_t^k || v_t^k), \tilde{f}_{t+1} = Att(p_t) \tag{7}$$

It is worth noting that $Att(\cdot)$ in the attention module should be selected carefully. Considering the characteristics of memories p_t, which contains the resolved spatial features at the pixel level with various semantics in different channels, the spatial-channel attentions should be used for a comprehensive capture of significant correlations, such as CBAM [44], Triplet [45], and CoT [46]. We make a comparative study for determining the optimal attention layer in Section IV.B.

As for the causality module, we use Transfer Entropy (TE) to quantify the spatial-temporal causality (including the intensity and direction) among all grids of a certain feature map, which acts on memories p_t to make causal inferences for causal future fc_{t+1} as (8) analogous to the attention matrix (the yellow shadow of Fig. 2(b)).

$$fc_{t+1} = \mathcal{C}(h_t^k, v_t^k, fa_{t+1}) = TE \times p_t \tag{8}$$

Here, TE is constructed as an $N \times N$ matrix, where $N = H \times W$ is the size of the feature map and H/W represents the height/width. Finally, the correlated future in (6) and causal future in (8) are added together as (5) to complement each other.

C. The Implementation of TE

In (8), TE is an information-theoretic measurement of causality proposed by [47], the calculation of which depends on both the historical and predicted (future) information. Continuing the example in the Introduction, taking the spatial-temporal features ($x^{(i)}$ and $x^{(j)}$) on two different grids ($i, j = 1{:}\,N$) of a feature map as an example, the causal relationship quantified by TE between them can be measured as (9),

$$te_{j\to i} = \sum \mathcal{P}\left(x_{t+1}^{(i)}, x_{1:t}^{(i)}, x_{1:t}^{(j)}\right) \log \frac{\mathcal{P}(x_{t+1}^{(i)} | x_{1:t}^{(i)}, x_{1:t}^{(j)})}{\mathcal{P}(x_{t+1}^{(i)} | x_{1:t}^{(i)})} \tag{9}$$

where $x_{1:t}^{(i)}$ is the historical feature on grid i, $x_{1:t}^{(j)}$ represents the historical feature on grid j to be investigated, and $x_{t+1}^{(i)}$ represents the future feature to be predicted. $\mathcal{P}(\cdot,\cdot,\cdot)$ is the joint probability and $\mathcal{P}(\cdot\,|\,\cdot)$ is the conditional probability. According to this formula, TE can be understood intuitively as the variations of the information entropy of $x_{t+1}^{(i)}$ when $x_{1:t}^{(j)}$ is known or not. Extensive research indicates that TE does not need to assume the form of the causal relationship between grids, which is suitable for the long-time series analysis of nonlinear systems [48].

To circumvent the probability calculation in neural networks, we rewrite (9) as conditional mutual information and divided into a combination of several simple terms in (10).

$$\begin{aligned} te_{j\to i} &= I(x_{t+1}^{(i)}; x_{1:t}^{(j)} | x_{1:t}^{(i)}) \\ &= \mathcal{H}\left(x_{t+1}^{(i)}, x_{1:t}^{(i)}\right) + \mathcal{H}\left(x_{1:t}^{(j)}, x_{1:t}^{(i)}\right) \\ &\quad -\mathcal{H}\left(x_{t+1}^{(i)}, x_{1:t}^{(j)}, x_{1:t}^{(i)}\right) - \mathcal{H}(x_{1:t}^{(i)}) \end{aligned} \tag{10}$$

where $I(\cdot;\cdot\,|\,\cdot)$ represents the conditional mutual information and $\mathcal{H}(\cdot)$ ($\mathcal{H}(\cdot,\cdot)$ and $\mathcal{H}(\cdot,\cdot,\cdot)$) represents the (joint) entropy. After nested traversing i and j, we can get a cause-effect matrix $TE \in \mathbb{R}^{N\times N}$ as shown in (11), which is similar to the attention matrix.

$$TE = \begin{bmatrix} te_{1\to1} & te_{2\to1} & te_{3\to1} & \cdots & te_{N\to1} \\ te_{1\to2} & te_{2\to2} & te_{3\to2} & \cdots & te_{N\to2} \\ te_{1\to3} & te_{2\to3} & te_{3\to3} & \cdots & te_{N\to3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ te_{1\to N} & te_{2\to N} & te_{3\to N} & \cdots & te_{N\to N} \end{bmatrix} \tag{11}$$

Following the such generic example, we independently learn and mine the spatial-temporal causal relationships between individual grids of the correlated future $fa_{t+1} \in \mathbb{R}^{C\times H\times W}$ and the previous memories $p_t \in \mathbb{R}^{C\times H\times W}$ (C represents channel). Specifically, we reshape both fa_{t+1} and p_t

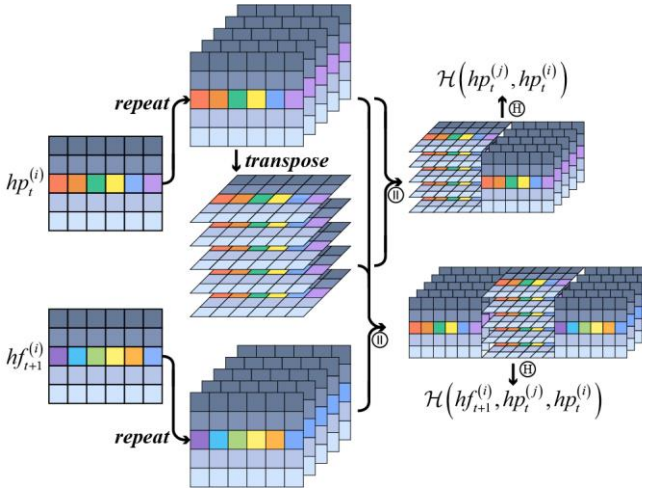

Fig. 3. Our proposed space-time exchange strategy for accelerating the calculation of TE in CAU. It requires replication of tensors and ordered concatenation.

into $\mathbb{R}^{N\times C}$ ($N = H \times W$). For a memory feature $p_t^{(i)} \in \mathbb{R}^{1\times C}$ on an individual grid $i = 1{:}N$ and the corresponding correlated future state $fa_{t+1}^{(i)} \in \mathbb{R}^{1\times C}$, we use TE to identify and quantify the causal contribution of states $p_t^{(j)} \in \mathbb{R}^{1\times C}$ of other grids $j = 1{:}N$ on the inference of $(p_t^{(i)}, p_t^{(j)}) \to fa_{t+1}^{(i)}$ (Here, we do not exclude i from the value range of j, because the causal relationship of itself cannot be ignored). We mark this process as $te_{j\to i} = I\left(fa_{t+1}^{(i)}; p_t^{(j)} \middle| p_t^{(i)}\right)$ like conditional mutual information in (10).

Before calculating $te_{j\to i}$, we set two transformation weights $W_p, W_f \in \mathbb{R}^{c\times\beta}$ (β is the hyper-parameter of hidden dimension, as shown in the two orange boxes in the yellow shadow of Fig. 2(b)) to precisely assess the (joint) entropy of the historical state and correlated future state with [0,1]-normalization as (12),

$$hp_t^{(i)} = \sigma(W_p \times p_t^{(i)}), hf_{t+1}^{(i)} = \sigma(W_f \times fa_{t+1}^{(i)}) \quad (12)$$

where σ is the sigmoid activation function. $hp_t^{(i)}$ and $hf_{t+1}^{(i)}$ are the normalized entropy. We make a comparative study for determining the optimal hidden dimension β in Section IV.B. Subsequently, $te_{j\to i}$ can be obtained by (13) analogous to (10) and forms the cause-effect matrix TE. This process is shown in the gray boxes of the yellow shadow in Fig. 2(b).

$$\begin{aligned} te_{j\to i} &= I\left(hf_{t+1}^{(i)}; hp_t^{(j)} \middle| hp_t^{(i)}\right) \\ &= \mathcal{H}\left(hf_{t+1}^{(i)}, hp_t^{(i)}\right) + \mathcal{H}\left(hp_t^{(j)}, hp_t^{(i)}\right) \\ &\quad -\mathcal{H}\left(hf_{t+1}^{(i)}, hp_t^{(j)}, hp_t^{(i)}\right) - \mathcal{H}\left(hp_t^{(i)}\right) \end{aligned} \quad (13)$$

Furthermore, we take advantage of the asymmetry of TE to identify the both intensity and direction of causal relationships. On the one hand, the larger $te_{j\to i}$, the greater the effect of grid j on i, the more reliable causality. On the other hand, if $te_{j\to i} > te_{i\to j}$, the feature of grid j is the effect while the feature of grid i is the cause. We then use (14) to remove the "effect" and retain the "cause",

$$\widetilde{TE} = max(TE^T - TE + diag(TE), 0) \quad (14)$$

where $\cdot^T$ represents the transpose and $diag(\cdot)$ represents the diagonal to supplement its own causality (e.g., $te_{1\to 1}$, $te_{2\to 2}$, etc.). After such filter, we normalize the cause-effect relationships by a mask-Softmax operator as (15).

$$\widetilde{TE}_{ms} = mask\_softmax\left(\widetilde{TE}_{i,:}\right) = \frac{e^{te_{j\to i}}}{\sum_{l=1}^{N} e^{te_{j\to i}}} \quad (i,j = 1{:}N, if\ te_{j\to i} \neq 0) \quad (15)$$

We practically use the normalized causality map $\widetilde{TE}_{ms}$ to augment the historical memory p_t to generate the causal feature in (8).

D. The Accelerated Calculation of TE

The calculation of TE requires two levels of loops (i and j), which is a time-consuming process. We use space-time exchange for acceleration. After in-depth analysis, two ingredients of $te_{j\to i}$ are more difficult to compute: $\mathcal{H}\left(hp_t^{(i)}, hp_t^{(j)}\right)$ and $\mathcal{H}\left(hf_{t+1}^{(i)}, hp_t^{(j)}, hp_t^{(i)}\right)$, because they are related to the traverse of i and j concurrently. Therefore, we design a replication-cascade strategy to construct the traverse as shown in Fig. 3.

As shown in this figure, the historical state and the correlated future are first replicated along the "Z-"axis, meanwhile the replicated historical state also needs to transpose the second and third dimensions, which is to construct a traversal cascade for all different positions. Then, we cascade them in the required order and calculate their joint entropies respectively. Such operation can construct the traverse of n and l simultaneously. The time-consuming is reduced from $\mathcal{O}(N^2)$ to $\mathcal{O}(N)$ according to mathematical analysis ($N = H \times W$ represents the size of the feature map). We also perform a validation experiment to monitor the elapsed time of these two strategies when calculating $\widetilde{TE}_{ms}$ under $N = 16 \times 16$ with our computing resource mentioned in Section IV.A. The average time-consuming for the former is 6.017 (ms) and that for the latter is 0.030 (ms), which is nearly a 200-fold increase (two orders of magnitude of acceleration ratio).

E. A Simple General Model for STSF Problem based on CAU

As mentioned above, CAU can be easily combined with existing ConvRNN cells for solving the STSF problem. We

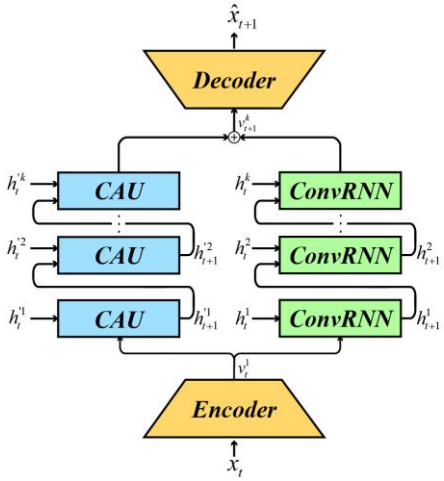

Fig. 4. The detailed structure of the encoder and decoder of our proposed model, which both consist of stage blocks and transition blocks. CNR and TCNR are our defined integrated components.

Table I

DESCRIPTION OF OUR CHOSEN DATASET AND EXPERIMENTAL SETTINGS FOR THEM. THE TRAINING AND TESTING PARADIGMS DENOTE THE NUMBER OF THE FRAME AS THE INPUT AND THE OUTPUT DURING TRAINING AND TESTING.

Dataset	Resolution	Training Paradigm	Testing Paradigm
Moving MNIST [43]	$1 \times 64 \times 64$	$10 \rightarrow 10$	$10 \rightarrow 10$
KTH [44]	$1 \times 64 \times 64$	$10 \rightarrow 10$	$10 \rightarrow 40$
BAIR [45]	$3 \times 64 \times 64$	$10 \rightarrow 10$	$10 \rightarrow 40$
WeatherBench [46]	$1 \times 32 \times 64$	$24 \rightarrow 24$	$24 \rightarrow 96$

build a simple general model based on it according to (3), as in Fig. 4.

We use an encoder and a decoder respectively to resolve and restore the spatial-temporal features hidden in the input sequence in this model. Between the encoder and decoder, the stacked CAU and ConvRNN cells are constructed in parallel. Analogous to ConvRNN, each layer of CAU receives the output of the previous layer and maintains the individual historical memories. The output of CAU is added to the output of ConvRNNs and subsequently fed to the decoder to make forecasting end-to-end. In this model, CAU assists in complementing the causal inference ability while ConvRNNs are performing memory updating. The selection of ConvRNN cells is important, and we perform experiments to verify the performance of CAU combined with different ConvRNN cells (See Section IV.D).

As for the encoder and decoder, we construct the fully-convolution networks respectively as shown in Fig. 5, which are both composed of two integrated blocks, that is, the stage block and transition block. The stage block consists of two CNRs (TCNRs) with different convolution kernel sizes, which resolve (restore) the spatial information by expanding (shrinking) the channel. The transition block is also constructed by CNR or TCNR, but the stride of the convolution layer is set as 2, which compresses the redundant information by reducing the feature maps and restores image details end-to-end by enlarging the feature maps in the trainable way.

In general, using the residual connections or some pooling/upsampling can effectively improve the performance of network, such as bridging the short-cut in the stage block, using the pooling/upsampling layers in the transition block. However, after extensive experiments, we find these two modifications are not suitable for improving the performance (See Section IV.D). We summarize that the short-distance residual connections produce the constant mapping easily, obviously preventing the features from evolving with time lines and keeping them align with historical features. In addition, the untrainable pooling/upsampling may violently discard the crucial details or introduce the redundant noise in the feature maps.

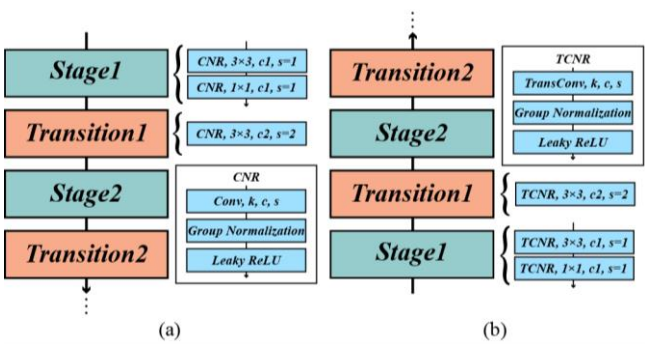

Fig. 5. The architecture of our proposed model. CAU is constructed parallel to vanilla ConvRNN cells in a plug-and-play way, which can complement the ability of causal mining. CAU can be stacked in multiple layers.

Subsequently, we tune the model configuration to the optimal performance (See Section IV.B) and solve the STSF problem in a common scenario (Moving MNIST dataset), video predictions (KTH and BAIR datasets), and weather forecasting (WeatherBench dataset) to comprehensively evaluate its performance (See Section IV.C). In addition, we also perform some ablation and sensitive studies for our model (See Section IV.D).

IV. EXPERIMENTS

A. Experiment Schemes

Datasets We select 4 different datasets from different scenarios to evaluate the performance of CAU. ① **Moving MNIST** [49], the most widely used benchmark dataset for the STSF problem, which is an ideal scenario of handwritten number twisting and shifting. ② **KTH** [50], a human performing video dataset containing 6 different actions, including walking, jogging, running, boxing, hand waving, and clapping. ③ **BAIR** [51], an object-moving video dataset pushed by a robotic arm. ④ **WeatherBench** [52], a global hourly weather reanalysis data from 1979 to 2018, from which we collect the sub-sets of air temperature and geopotential. The image size, training paradigm, and testing paradigm are shown in Table I.

Model Settings and Training We conduct all the subsequent experiments on a server with a GPU of NVidia RTX 3090, and all models are optimized with Adam [47] for MSE (mean square errors). In addition, we have first carried out a series joint tuning experiments to determine optimal major hyper-parameters under our computing resource (See Section IV.B). To summarize, Triplet is selected as the attention module in CAU, the block number in encoder/decoder is 3, the CAU number is 3, and the hidden dimension for the calculation of TE is 48. For a more robust training effect, we set a "warm-up" phase in the early training epoch. In this phase, the recurrent training of model just uses the standard label rather than the previous time step's prediction, and the "warm-up" length is equal to the input length as shown in Table I. After the "warm-up" phase, the model performs the formal iterative prediction. The "warm-up" phase can make the model load more correct memories and guide a precise/rapid direction of parameter learning in the early training epoch. We set the "warm-up" phase as 50 epochs for the subsequent experiments.

Metrics We select 3 different metrics to evaluate model performance, including MSE (lower is better), $PSNR$ (peak signal-to-noise ratio, larger is better), and $SSIM$ [53] (structural

similarity, larger is better). These three metrics focus on not only the prediction errors at the pixel level but the human-eye perception of the predictions. The precise equations of these three metrics can be referred to as (16) to (18).

$$MSE = \frac{1}{\kappa}\sum_{s=t+1}^{t+\kappa}\left(\hat{\mathcal{X}}_s^{(i)} - \mathcal{X}_s^{(i)}\right)^2, i = 1{:}N \tag{16}$$

$$PSNR = 10 \cdot log_{10}\left(\frac{MAX^2}{MSE}\right) \tag{17}$$

$$\begin{cases} luminance = \frac{2\mu_{\hat{\mathcal{X}}}^2\mu_{\mathcal{X}}^2 + c_1}{\mu_{\hat{\mathcal{X}}}^2 + \mu_{\mathcal{X}}^2 + c_1} \\ contrast = \frac{2\sigma_{\hat{\mathcal{X}}\mathcal{X}} + c_2}{\sigma_{\hat{\mathcal{X}}}^2 + \sigma_{\mathcal{X}}^2 + c_2} \\ structure = \frac{\sigma_{\hat{\mathcal{X}}\mathcal{X}} + c_3}{\sigma_{\hat{\mathcal{X}}}\sigma_{\mathcal{X}} + c_3} \\ SSIM = luminance^a \cdot contrast^b \cdot structure^c \end{cases} \tag{18}$$

where $\hat{\mathcal{X}}_s^{(i)}$ and $\mathcal{X}_s^{(i)}$ are the ground truth and prediction result on the grid i in the predicted sequence respectively, κ represents the sequence length. In $PSNR$, MAX is set as 255. In $SSIM$, $\mu_{\hat{\mathcal{X}}}$ ($\mu_{\mathcal{X}}$) is the average for $\hat{\mathcal{X}}$ ($\mathcal{X}$), and $\sigma_{\hat{\mathcal{X}}}$ ($\sigma_{\mathcal{X}}$) is the corresponding standard deviation. $\sigma_{\hat{\mathcal{X}}\mathcal{X}}$ represents the covariance, $a = b = c = 1$ for fair measurement of every ingredient of $SSIM$, c_1, c_2, and c_3 are all trivial values for preventing the denominator from being 0.

B. Determination of the Optimal Structure

The interior structure of CAU consists of many adjustable modules that influence the model performance. Here, we majorly divide them into two categories according to different scales: The structure-level macro design and the parameter-level micro design. The former contains the choice of attention module in CAU and the depth of encoder/decoder, and the latter contains the hidden dimension of the causality module and the stack number of CAU in the entire network. We first determine the optimal combination for the macro design and then tune the parameters of the micro design with the help of the Moving MNIST dataset. Note that other parameters also affect the model performance, such as the convolution kernel (channel and size). We omit the tuning process of these parameters because these are not the focus of this paper. In addition, we have adjusted them to the optimal in the subsequent experiments.

Macro Design For the attention module, we have selected 3 candidates according to the review of [54]: CBAM [44], X-Linear [55], Triplet [45], Coordinate [56], and CoT [46]. They are all plug-and-play spatial-channel attention techniques and have different characteristics. CBAM calculates the spatial and channel attentions individually and is the most widely used technique. X-Linear exploits the spatial-channel-wise bilinear attention distributions to capture the 2nd (or even infinity) order interactions between the multi-modal features. Triplet focuses on the cross-domain interactions between spatial and channel levels and aggregates their information together. Coordinate incorporates the position encoding of feature maps into the capture of spatial-channel attention. CoT capitalizes on the contextual information among input keys to guide the learning of dynamic attention matrix and thus strengthens the capacity of visual representation. These five types of attention mechanisms can cover most other options according to their motivations and implementations, and we select one from them for CAU.

For the depth of the encoder/decoder, we can deepen the model by increasing the number of stage modules before the transition module. (To ensure the size of the encoded features is not too small, we fix the number of transition modules to 2). Considering the above factors, we carry out an experiment to evaluate the joint effects of the attention module and model depth. Here, the CAU number is 1 and the hidden dimension is 36. The result is shown in Table II.

The effects of different attention modules are not much different with fluctuation of no more than 1 under the same block number, among which Triplet exhibits slightly high performance. This is because Triplet stresses the discriminative interactions of features in different angles, which is useful for capturing the rotating or shifting of the object during temporal evolutions. CoT also possesses a comparable performance, because it emphasizes the neighborhood interactions, strengthening the localized historical memory correlations and

Table II

THE JOINT EFFECTS ON PERFORMANCE WITH MULTIPLE MACRO DESIGNS, WHICH CONTAIN DIFFERENT ATTENTION MODULES IN CAU AND BLOCK NUMBERS OF ENCODER/DECODER. (THE CAU NUMBER IS 1 AND THE HIDDEN DIMENSION IS 36). **BOLD NUMBERS** REPRESENT THE BEST PERFORMANCE.

Attention Module	Block Number	Metrics	
		MSE	SSIM
CBAM [39]	1	35.4	0.917
	2	34.9	0.919
	3	34.4	0.920
Triplet [40]	1	34.6	0.919
	2	33.9	0.921
	3	**33.4**	**0.922**
Coordinate [50]	1	35.1	0.917
	2	34.8	0.919
	3	33.8	0.921
X-Linear [56]	1	35.2	0.916
	2	34.6	0.918
	3	34.2	0.918
CoT [46]	1	34.5	0.917
	2	34.0	0.919
	3	33.6	0.920

Table III

THE JOINT EFFECTS ON PERFORMANCE WITH MULTIPLE MICRO DESIGNS, WHICH CONTAIN DIFFERENT CAU NUMBERS AND HIDDEN DIMENSIONS FOR ENTROPY CALCULATION IN TE OF CAU. (THE ATTENTION MODULE IS TA AND THE BLOCK NUMBER IS 3). **BOLD NUMBERS** REPRESENT THE BEST PERFORMANCE.

CAU Number	Hidden Dimension β	Metrics	
		MSE	SSIM
1	24	35.7	0.915
	36	33.4	0.922
	48	31.7	0.924
2	24	34.9	0.918
	36	30.6	0.926
	48	28.5	0.931
3	24	33.1	0.922
	36	29.6	0.929
	48	**26.7**	**0.939**

spatial-temporal dynamic evolutions. On the other hand, the larger block number has a significant performance improvement. Because a deeper structure favors resolving and restoring the spatial features.

According to the best performance, we use Triplet as the attention module in CAU and the block number in the encoder/decoder is set as 3 in the subsequent experiments.

Micro Design The number of stacked CAUs affects the efficiency of memory propagation, and the hidden dimension β is closely related to the accuracy of entropy calculation in TE. These are two critical parameters. We design an experiment to determine the optimal combination of them. The result is shown in Table III.

Overall, more CAUs bring an obvious performance gain. But stacking CAU also induces a large computational load, and the improvement rate gets slower with the CAU number from 1 to 3. Meanwhile, increasing the hidden dimension β undoubtedly promotes the points of MSE and $SSIM$, which makes the calculation of entropy more precise. The parameter settings for this experiment are the widest adjustable ranges under our computing resources. In the future, more complex parameter determinations will be performed on larger-scale high-performance devices.

According to the best performance, we set the CAU number as 3 and the hidden dimension β as 48 in the subsequent experiments.

C. Comparisons with Other ConvRNN-based STSF Models

In this section, we compare the results of our proposed model and other representative STSF models on the four chosen datasets. Specifically, the comparisons of quantitative metrics are depicted in Table IV to Table VII, and the comparisons of visual quality are presented in Fig. 6 to Fig. 9. Note that some results of other models are either collected from the official papers or remade by the official codes.

Moving MNIST Fig. 6 illustrates the forecasting results, where our model obviously outperforms the other methods with sharper edges and more accurate placement of digits. Specifically, the visual qualities of these models are similar before $T = 20$, while the prediction qualities of PredRNN++ and E3D-LSTM drop steadily after $T = 24$. In addition, there is little difference between our model and MAU in human-eye perception, but our model has improvements numerically as shown in Table IV, which are not easy to detect.

For the comprehensive comparison, we also add the official records of other STSF models on this dataset. From Table IV, though our model has not achieved the best performance, it can still compensate for the causal mining ability of the basic ConvLSTM units, bringing the performance of it up to first-tier level. In addition, our model is not a sequence-input model and has no gated structures, which means it has fewer trainable parameters and easy convergence (See Section IV.D), saving a lot of time and computation resources. It is worth noting that our model does not achieve the best performance on this dataset, which are SimVP [26] (23.8@ MSE, 0.948@ $SSIM$) and MogaNet [25] (15.67@MSE, 0.966@$SSIM$). Instead of using ConvRNN-based structure, they both utilize the autoregression to predict the future directly via the conv-skeleton or the transformer-style models. This may be a better choice trend for

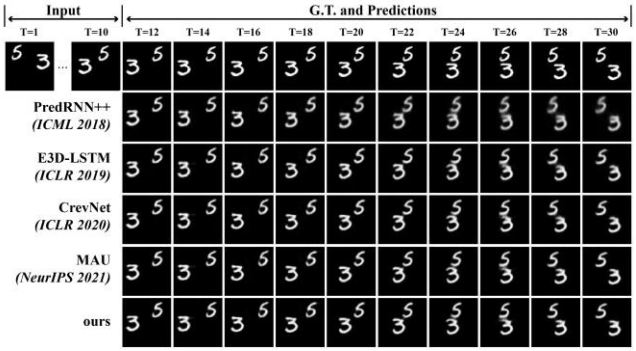

Fig. 6. Predictions of different methods on the Moving MNIST dataset ($10 \rightarrow 30$).

prediction accuracy. However, our model emphasizes more on temporal memory transfer of correlation and causality, which can effectively help bring the performance of multiple ConvRNN modules up to first-tier level, and is more in line with the laws of development of things in the physical/natural world and the human understanding of the spatial-temporal evolution of objects.

KTH Our model can achieve excellent performance on the KTH dataset as shown in Fig. 7. Previous models (PredRNN++ and E3D-LSTM) can only forecast the rough position and boundary of the standing man, but our model can predict finer and clearer movements, such as arm waving and clapping. Compared with STAM (the upgrade of MAU), our model has a slight improvement, especially after $T = 35$, exhibitting more precise description of motion details.

Table V shows the numerical comparisons between our model and other models. Our model gains improvements on these two measurements both, especially when extending the forecast lead time from 10 to 30. On the other hand, the performance of our model can maintain a high level with the forecast lead time increases. We speculate that our model is less affected by accumulated errors during forecasting iteration. Because it receives one-frame data as input, which contains fewer errors. While most other models receive sequential data as input, which involves more uncertainties. Besides, the performance of our model is comparable to the conv-skeleton model SimVP [26], the 20-step performance of which achieves 33.72@$PSNR$ and 0.905@$SSIM$, and the 40-step performance of which achieves 32.93@$PSNR$ and 0.886@$SSIM$. This also reflects the high degree of consistency and generalization of our model in iterated forecasts.

BAIR This dataset contains more sophisticated scenarios and actions, but our model can also forecast future frames with the better visual quality compared with other models as shown in Fig. 8. It can be seen from two aspects. On the one hand, the background (objects) of the predictions maintains a high consistency, which shows little change as the lead time increases. On the other hand, although the robotic arm is blurred to some extent, its position and size can be predicted more accurately compared to other models.

Table VI shows the quality measurements. It indicates that our model can achieve comparable MSE and $SSIM$ scores, which exhibits the superiority of our proposed model.

WeatherBench Fig. 9 shows the prediction results of temperature and geopotential respectively. As for this dataset,

Table IV

QUANTITATIVE RESULTS OF DIFFERENT MODELS ON THE MOVING MNIST DATASET. THE METRICS ARE AVERAGED OVER THE PREDICTED FRAMES (10→10). **BOLD NUMBERS** REPRESENT THE BEST PERFORMANCE, UNDERLINED NUMBERS REPRESENT THE SECOND-BEST PERFORMANCE.

Model	*MSE*	*SSIM*
ConvLSTM [11]	103.3	0.707
PredRNN [20]	55.8	0.867
MIM [22]	44.2	0.910
PredRNN++ [21]	46.5	0.898
E3D-LSTM [23]	41.3	0.910
SA-ConvLSTM [29]	43.9	0.913
CrevNet [51]	38.5	0.928
MAU [14]	29.5	0.931
PhyDNet [24]	**24.4**	**0.947**
CAU (ours)	26.7	0.939

Table V

THE SAME WITH Table IV, BUR FOR THE KTH DATASET.

Model	10 → 20		10 → 40	
	PSNR	*SSIM*	*PSNR*	*SSIM*
ConvLSTM [11]	23.6	0.712	22.9	0.639
SAVP [52]	25.4	0.746	24.0	0.701
PredRNN++ [21]	28.5	0.865	25.2	0.741
E3D-LSTM [23]	29.3	0.879	27.2	0.810
SRVP [53]	30.1	0.885	28.6	0.816
WAM [54]	29.9	0.893	27.5	0.854
STAE [55]	29.9	0.899	27.9	0.859
STAM [15]	**30.5**	**0.929**	28.9	0.906
CAU (ours)	30.4	0.927	**29.1**	**0.911**

all of the models can achieve a high-level performance, which is because the weather changes in the hour scale are very few.

Numerically, our model still leads the way in this dataset. Table VII displays the comparison results. From 24-hour to 96-hour forecasts, our model has achieved the best scores on all metrics, which shows that it has a good potential for inference of weather evolution.

Summary of Four Datasets For the 4 chosen datasets, our model can effectively improve the performance of existing ConvRNN cells in a plug-and-play way with outstanding quantitative scores and visual quality, reaching the SOTA level, especially in the long-term forecasts. It implies that our proposed CAU is an effective tool for mining causality hidden in long-term spatial-temporal evolutions, which is exactly the lack of vanilla memory units.

D. Performance Evaluations

Training Process Analysis During the training process of the 4 chosen datasets, an interesting phenomenon occurs: Our model converges faster than other models on the testing set. For example, Fig. 10 shows the performance curves of 4 different models as the training epochs increase on Moving MNIST and WeatherBench (T300) datasets, which is the average result of 20-time training. The blue lines belong to our model, which has a larger slope than others, especially in weather forecasting. This indicates that time-/resource-consuming will be saved.

We think this is majorly due to the non-gated structure of CAU compared to traditional RNN. During the training of RNN, the gradient vanishing (or explosion) easily shows up, especially in a non-stationary series and long-term propagation. The gated structures may exacerbate this phenomenon by generating a quite strict gated mask, because it is shared by all frames and easily ignores location-variant anomalies, such as rotation and scaling, etc. Therefore, the performance curves of gated-based RNN models (PredRNN++, E3D-LSTM, MAU) are gentler. CAU has no gated modules, the training of which is easier to converge.

Table VI

THE SAME WITH Table IV, BUR FOR THE BAIR DATASET.

Model	10 → 20		10 → 40	
	PSNR	*SSIM*	*PSNR*	*SSIM*
SAVP [52]	20.5	0.844	18.4	0.795
SVG [43]	21.2	0.857	19.0	0.816
PredRNN++ [21]	21.0	0.849	18.6	0.803
E3D-LSTM [23]	21.1	0.851	19.1	0.814
SRVP [53]	23.7	0.867	19.6	0.820
WAM [54]	**25.4**	0.881	**21.0**	0.844
CAU (ours)	25.1	**0.884**	20.9	**0.846**

Table VII

THE SAME WITH Table IV, BUR FOR THE WEATHERBENCH DATASET (10→10).

Model	Z500 (*RMSE*)	T850 (*RMSE*)
PredRNN [20]	331.2	1.49
MIM [22]	323.6	1.45
E3D-LSTM [23]	308.6	1.45
SA-ConvLSTM [29]	314.2	1.47
CrevNet [51]	283.7	1.34
MAU [14]	253.9	1.29
CAU (ours)	**237.0**	**1.25**

Ablation and Sensitive Studies To validate the effectiveness of the sub-modules (i.e., attention and causality modules) and the robustness of the model structure, we carry out the following four experiments in this section. Note that for a fair comparison, we uniformly use the Moving MNIST dataset as the STSF scenarios, and the other hyper-parameters are set the same. Table VIII shows the results.

- **Exp.1** Remove one of the sub-modules of CAU.
- **Exp.2** Swap the two sub-modules of CAU.
- **Exp.3** Use short-cut residual connections in stage block and pooling/upsampling layer in transition block
- **Exp.4** Add a gate module in CAU (Appending a vanilla RNN above two modules in CAU).
- **Exp.5** Exchange other spatial-temporal cells parallel to CAU.

As for Exp.1, we find that the absence of sub-modules leads to a severe performance drop, especially when removing the causal module, the *MSE* score has dropped 6 points. This indicates the effectiveness of our proposed causality module and particularly the combination of attention and causality.

As for Exp.2, the reverse of attention and causality modules causes performance loss to a certain extent. We think this is due to the failure of the attention module. When swapping the attention and causality modules, CAU will first perform the causal inference and then compute the correlations between different grids based on the causality-augmented features. However, the causality-augmented features have already filtered the spurious associations (with a mask-Softmax operator), which means the attention modules make a repetitive and ineffective contribution. So our proposed formalization is a better choice for the order of attention and causality modules.

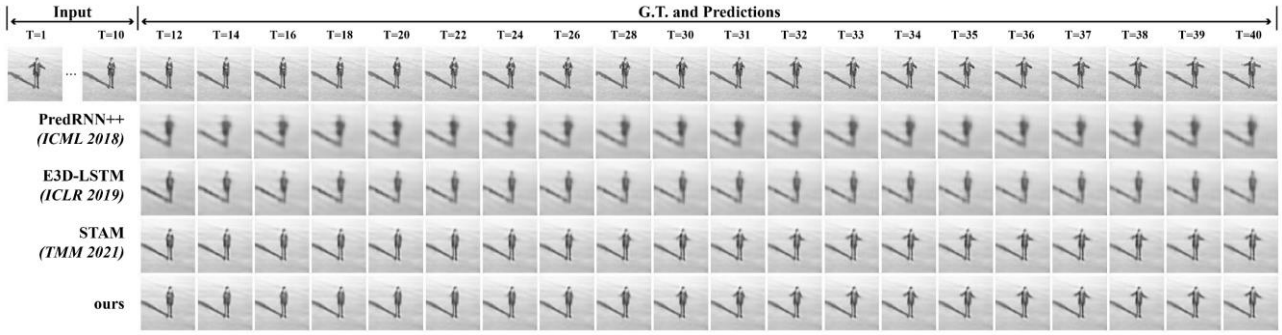

Fig. 7. Predictions of different methods on the KTH dataset (10 → 40).

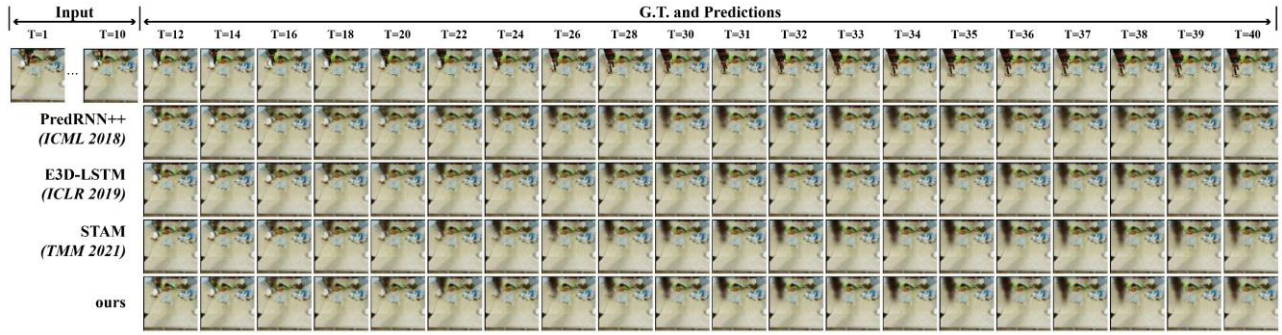

Fig. 8. Predictions of different methods on the BAIR dataset (10 → 40).

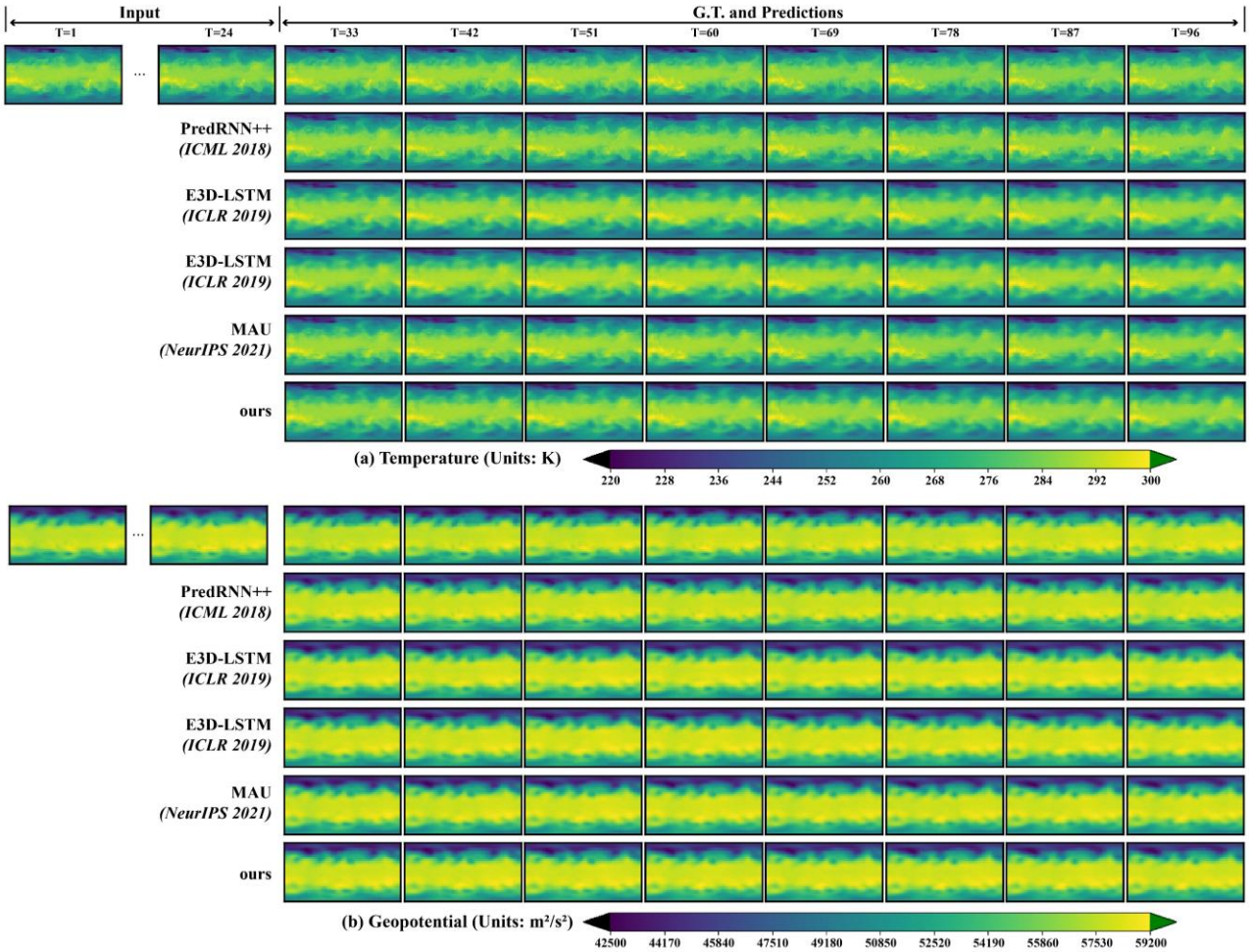

Fig. 9. Predictions of different methods on the WeatherBench dataset (24 → 96). The above sub-figure is for temperature, and the below sub-figure is for geopotential.

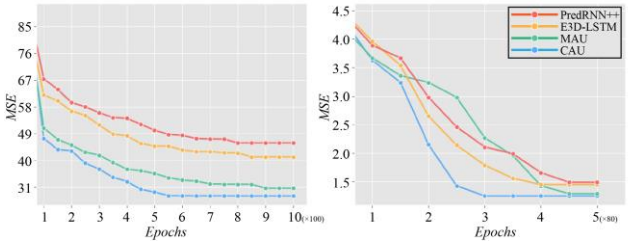

Fig. 10. The comparisons of performance on the testing set during training between our model and others. The left sub-figure is for the Moving MNIST dataset, and the right sub-figure is for the WeatherBench dataset (temperature).

As for Exp.3, these two structures both influence the model performance. Specifically, using pooling/upsampling layers in the transition block has a more severe impact on STSF models. The short-cut residual connections tend to homogenize predictions, which is not good for distinguishing variations over time. Pooling/upsampling layers with no learnable parameters may discard important spatial information and introduce redundant noise, which lead to large uncertainties in resolving and restoring spatial features. It provides a solid suggestion for the construction of the encoder/decoder of future STSF models. In addition, in order to fully utilize the advantages of residual connections, the feature maps in the encoder can be "remotely" connected to the decoder, which eliminates the constant mapping and makes the STSF model recognize more spatial-temporally discriminative features to improve performance, like [14].

As for Exp.4, the gated structures are serious handicaps for CAU from the experiment results. This is major because the mask-Softmax operator in CAU has been equivalent to a gated structure from the aspect of the memory filter. If other gated structures are added, the output of the entire CAU will become extremely harsh (i.e., the memory filtering will become very strict), which is very likely that the gated unit is all 0. In a word, the gated modules are not suitable for integration in CAU.

As for Exp.5, the exchange of other spatial-temporal cells indeed helps improve the performance of CAU. The paralleled spatial-temporal cells complement the insufficiency of the CAU for memory propagation, so the better the long-term memory processing, the better the performance. It can be seen from the table that the MIM cell is a better choice for CAU, which is good at capturing the changes between memories, maintaining both instantaneous and tendency. On the other hand, CAU, as a plug-and-play unit, has strong robustness and can effectively improve the performance of existing models simply.

Visualization of Causal Maps To interpret what CAU has learned, we choose a typical scenario in the Moving MNIST dataset and visualize the learned causal maps of the crucial pixel as shown in Fig. 11. In this figure, the big image in every sub-figure represents the prediction results of CAU, and two small images are the visualizations of causal maps of the corresponding pixel in colored boxes (marked as orange and green pixels respectively) on the big image, which are both the "head" of each digit.

Although the position characteristics of these two pixels are similar, their causal maps are quite distinctive. As for the pixel of digit "2" (the causal maps of which are the orange small images), the sensitive regions of the causal maps are usually at the tail of digit "2" and merely located at the head, such as in (a)-(d). While as for the pixel of digit "7" (the causal maps of which are the green small images), the significant regions just spread around it. The tail of digit "2" and the head of digit "7" will meet and mix together during the whole evolution as shown in (f)-(h). This shows that CAU can find the most critical regions of object evolution in the image and make prominent processing.

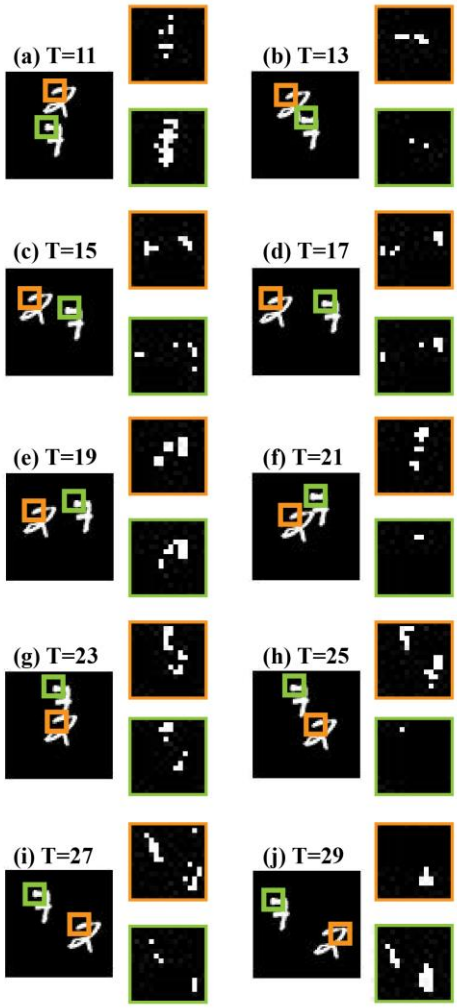

Fig. 11. The visualization of causal maps learned by CAU during prediction. The big image is the prediction result and the two small images are the causal maps in (a)-(j). Specifically, the causal maps with orange (green) borders are the causal maps of the pixel in the small orange (green) box on prediction results.

Meanwhile, the causal maps of these two pixels also imply that they are "monitoring" each other, especially when two digits are entangled. For example, there are two obvious highlighted regions in both causal maps in (a)-(g). But when the two digits are far apart, the individual causal maps tend to focus only on their own changes, such as in (h)-(j). It indicates the causal maps learned by CAU exhibit a good semantic segmentation ability, clearly separating the positions of the two digits. This is also the reason why the predictions of CAU are clearer and sharper.

Table VIII

ABLATION AND SENSITIVITY STUDIES ON THE MOVING MNIST DATASET FOR CAU.

No.	Modification	Metrics	
		MSE	*SSIM*
Exp.1	CAU w/o Attention module	30.1	0.930
	CAU w/o Causality module	35.4	0.914
Exp.2	Inverse CAU	34.7	0.917
Exp.3	Using residual connection in stage module	35.4	0.913
	Using pooling/upsampling in transition module	40.3	0.904
Exp.4	CAU + GRU	41.5	0.901
	CAU + LSTM	41.9	0.903
Exp.5	CAU || SA-ConvLSTM	26.7	0.936
	CAU || Causal-LSTM	26.5	0.938
	CAU || MIM	26.3	0.941

V. CONCLUSIONS AND FUTURE WORKS

In this paper, we propose a Causality Attention Unit (CAU) to improve the prediction accuracy and visual quality for the STSF problem. It contains two sequential attention and causality models, playing a significant role to compensate for the ability of ConvRNN-based models to mine causal relationships via the superiority of transfer entropy (TE). In addition, a time-space exchange strategy is designed to accelerate the calculation of TE, a simple general model for STSF problem based on CAU is constructed. To evaluate the model performance, we carry out multiple comparison experiments with other STSF models on four different public datasets (i.e., Moving MNIST, KTH, BAIR, and WeatherBench), accompanied with some ablation and sensitivity studies. The experiment results indicate that our proposed CAU-based model can effectively improve the quantitative measurements and visual qualities of existing ConvRNN cells (e.g., MIM, PredRNN++) in a plug-and-play way, reaching the SOTA level. The visualization of the learnt causal maps demonstrates that CAU is an outstanding tool for distinguishing pixel attribution and motion state in sophisticated entangled scenarios.

The demand for accurate forecasts is endless. In the future, we will make better use of the causal module in CAU, identifying key patterns (e.g., saliency map) interpretably. In addition, we will also continue to optimize and reduce the high memory occupation in TE's computation caused by the necessary tensor replication, and then explore the application of CAU in more research areas and datasets for a more comprehensive causal relationship mining.

REFERENCES

[1] X. Shi and D.-Y. Yeung, "Machine learning for spatiotemporal sequence forecasting: A survey," *ArXiv Prepr. ArXiv180806865*, 2018.

[2] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1683–1698, 2019.

[3] W. Gao *et al.*, "Digital retina: A way to make the city brain more efficient by visual coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4147–4161, 2021.

[4] X. Chen and W. Wang, "Uni-and-bi-directional video prediction via learning object-centric transformation," *IEEE Trans. Multimed.*, vol. 22, no. 6, pp. 1591–1604, 2019.

[5] H. Kalbkhani, M. G. Shayesteh, and N. Haghighat, "Adaptive lstar model for long-range variable bit rate video traffic prediction," *IEEE Trans. Multimed.*, vol. 19, no. 5, pp. 999–1014, 2016.

[6] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, 2015.

[7] B. Mu, B. Qin, and S. Yuan, "ENSO-ASC 1.0.0: ENSO deep learning forecast model with a multivariate air–sea coupler," *Geosci. Model Dev.*, vol. 14, no. 11, pp. 6977–6999, 2021.

[8] B. Mu, Y. Cui, S. Yuan, and B. Qin, "Simulation, Precursor Analysis and Targeted Observation Sensitive Area Identification for Two Types of ENSO using ENSO-MC v1. 0," *Geosci. Model Dev. Discuss.*, pp. 1–21, 2022.

[9] B. Mu, B. Qin, and S. Yuan, "ENSO-GTC: ENSO Deep Learning Forecast Model with a Global Spatial-Temporal Teleconnection Coupler," *J. Adv. Model. Earth Syst.*, p. e2022MS003132, 2022.

[10] B. Mu, B. Qin, S. Yuan, X. Wang, and Y. Chen, "PIRT: A Physics-informed Red Tide Deep Learning Forecast Model Considering Causal-inferred Predictors Selection," *IEEE Geosci. Remote Sens. Lett.*, 2023.

[11] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.

[12] X. Shi *et al.*, "Deep learning for precipitation nowcasting: A benchmark and a new model," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[13] A. Vaswani *et al.*, "Attention is all you need," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[14] Z. Chang *et al.*, "MAU: A Motion-Aware Unit for Video Prediction and Beyond," *Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.

[15] Z. Chang, X. Zhang, S. Wang, S. Ma, and W. Gao, "STAM: A SpatioTemporal Attention based Memory for Video Prediction," *IEEE Trans. Multimed.*, 2022.

[16] L. Tu, G. Lalwani, S. Gella, and H. He, "An empirical study on robustness to spurious correlations using pre-trained language models," *Trans. Assoc. Comput. Linguist.*, vol. 8, pp. 621–633, 2020.

[17] Z. Wang and A. Culotta, "Identifying spurious correlations for robust text classification," *ArXiv Prepr. ArXiv201002458*, 2020.

[18] G. Woo, C. Liu, D. Sahoo, A. Kumar, and S. Hoi, "CoST: Contrastive Learning of Disentangled Seasonal-Trend Representations for Time Series Forecasting," *ArXiv Prepr. ArXiv220201575*, 2022.

[19] J. Peters, D. Janzing, and B. Schölkopf, *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[20] Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 879–888.

[21] Y. Wang, Z. Gao, M. Long, J. Wang, and S. Y. Philip, "Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning," in *International Conference on Machine Learning*, PMLR, 2018, pp. 5123–5132.

[22] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang, and P. S. Yu, "Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9154–9162.

[23] Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3d lstm: A model for video prediction and beyond," in *International conference on learning representations*, 2018.

[24] V. L. Guen and N. Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11474–11484.

[25] S. Li *et al.*, "Efficient multi-order gated aggregation network," *ArXiv Prepr. ArXiv221103295*, 2022.

[26] C. Tan, Z. Gao, and S. Z. Li, "SimVP: Towards Simple yet Powerful Spatiotemporal Predictive Learning," *ArXiv Prepr. ArXiv221112509*, 2022.

[27] Z. Gao *et al.*, "Earthformer: Exploring space-time transformers for earth system forecasting," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 25390–25403, 2022.

[28] C. Bai, F. Sun, J. Zhang, Y. Song, and S. Chen, "Rainformer: Features extraction balanced network for radar-based precipitation nowcasting," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[29] Z. Lin, M. Li, Z. Zheng, Y. Cheng, and C. Yuan, "Self-attention convlstm for spatiotemporal prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 11531–11538.

[30] H. Guo, D. Zhang, L. Jiang, K.-W. Poon, and K. Lu, "ASTCN: An Attentive Spatial Temporal Convolutional Network for Flow Prediction," *IEEE Internet Things J.*, 2021.

[31] C. Tan, Z. Gao, S. Li, Y. Xu, and S. Z. Li, "Temporal attention unit: Towards efficient spatiotemporal predictive learning," *ArXiv Prepr. ArXiv220612126*, 2022.

[32] J. Li, X. Liu, W. Zhang, M. Zhang, J. Song, and N. Sebe, "Spatio-temporal attention networks for action recognition and detection," *IEEE Trans. Multimed.*, vol. 22, no. 11, pp. 2990–3001, 2020.

[33] D. Li, T. Yao, L.-Y. Duan, T. Mei, and Y. Rui, "Unified spatio-temporal attention networks for action recognition in videos," *IEEE Trans. Multimed.*, vol. 21, no. 2, pp. 416–428, 2018.

[34] W. Du, Y. Wang, and Y. Qiao, "Recurrent spatial-temporal attention network for action recognition in videos," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1347–1360, 2017.

[35] B. Yang, Z. Tu, D. F. Wong, F. Meng, L. S. Chao, and T. Zhang, "Modeling localness for self-attention networks," *ArXiv Prepr. ArXiv181010182*, 2018.

[36] M. Guo, Y. Zhang, and T. Liu, "Gaussian transformer: a lightweight approach for natural language inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 6489–6496.

[37] Y. Luo, J. Peng, and J. Ma, "When causal inference meets deep learning," *Nat. Mach. Intell.*, vol. 2, no. 8, pp. 426–427, 2020.

[38] Z. Lv, J. Li, C. Dong, and Z. Xu, "DeepSTF: A Deep Spatial–Temporal Forecast Model of Taxi Flow," *Comput. J.*, 2021.

[39] S. Liu, S. Dai, J. Sun, T. Mao, J. Zhao, and H. Zhang, "Multicomponent spatial-temporal graph attention convolution networks for traffic prediction with spatially sparse data," *Comput. Intell. Neurosci.*, vol. 2021, 2021.

[40] X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing, "Dags with no tears: Continuous optimization for structure learning," *Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[41] X. Lin, Y. Chen, G. Li, and Y. Yu, "A causal inference look at unsupervised video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 1620–1629.

[42] J. Chen, X. Wu, Y. Hu, and J. Luo, "Spatial-temporal causal inference for partial image-to-video adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 1027–1035.

[43] A. Gu, C. Gulcehre, T. Paine, M. Hoffman, and R. Pascanu, "Improving the gating mechanism of recurrent neural networks," in *International Conference on Machine Learning*, PMLR, 2020, pp. 3800–3809.

[44] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[45] D. Misra, T. Nalamada, A. U. Arasanipalai, and Q. Hou, "Rotate to attend: Convolutional triplet attention module," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3139–3148.

[46] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 1489–1500, 2022.

[47] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, p. 461, 2000.

[48] S. Yang, L. Ning, X. Cai, and M. Liu, "Dynamic Spatiotemporal Causality Analysis for Network Traffic Flow Based on Transfer Entropy and Sliding Window Approach," *J. Adv. Transp.*, vol. 2021, 2021.

[49] E. Denton and R. Fergus, "Stochastic video generation with a learned prior," in *International conference on machine learning*, PMLR, 2018, pp. 1174–1183.

[50] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, IEEE, 2004, pp. 32–36.

[51] F. Ebert, C. Finn, A. X. Lee, and S. Levine, "Self-Supervised Visual Planning with Temporal Skip Connections.," in *CoRL*, 2017, pp. 344–356.

[52] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, "WeatherBench: a benchmark data set for data-driven weather forecasting," *J. Adv. Model. Earth Syst.*, vol. 12, no. 11, p. e2020MS002203, 2020.

[53] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[54] M.-H. Guo *et al.*, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, pp. 1–38, 2022.

[55] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10971–10980.

[56] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 13713–13722.

Bo Qin received the B.S. degree and the Ph.D. degree in School of Software Engineering, Tongji University, Shanghai, China, in 2017 and 2023 respectively.

He is currently a Post-Doctoral Researcher with the Department of Atmospheric and Oceanic Sciences & Institude of Atmospheric Sciences, Fudan University, Shanghai, China. He is mainly engaged in building intelligent simulating/forecasting models with a certain degree of physical interpretability for multiple weather/climate phenomena by artificial intelligence.

Fanqing Meng received the B.S. degree in School of Software Engineering, Tongji University, Shanghai, China. He is currently pursing the Ph.D. degree in School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.

His current research interest focuses on the applications of computer vision as well as multimodal and transfer learning.

Shijin Yuan received the B.S. degree in mechanical and electronic engineering from Wuhan University of Technology, Hubei, China in 1997; the master of engineering degree in control theory and control engineering from Zhejiang University of Technology, Zhejiang, China in 2001; and the Ph.D. degree in computer software and theory from Fudan University, Shanghai, China in 2004.

She was an associate professor from 2012 to 2016 and is now a professor in school of software engineering, Tongji University, Shanghai, China since 2016. She is mainly engaged in building intelligent forecasting models with a certain degree of physical interpretability for multiple weather/climate phenomena by artificial intelligence.

Bin Mu received the B.S. degree in computer science and technology from Anhui University, Anhui, China in 1985; and the master of engineering degree in computer science and technology from Hefei University of Technology, Hefei, China in 1990.

From 2002 to 2003, he was a senior visiting scholar in school of computer science of New Brunswick University, Canada. From 2003 to 2004, he was an associate professor in school of computer and information of Hefei University of Technology, China. He was an associate professor from 2004 to 2010 and is now a professor in school of software engineering of Tongji University, China since 2010. He is mainly engaged in building intelligent forecasting models with a certain degree of physical interpretability for multiple weather/climate phenomena by artificial intelligence.