# Temporal Attention Unit: Towards Efficient Spatiotemporal Predictive Learning

Cheng Tan*, Zhangyang Gao*, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, Stan Z. Li†
AI Lab, Research Center for Industries of the Future, Westlake University
{tancheng,gaozhangyang,wulirong,xuyongjie,xiajun,lisiyuan,stan.zq.li}@westlake.edu.cn

## Abstract

*Spatiotemporal predictive learning aims to generate future frames by learning from historical frames. In this paper, we investigate existing methods and present a general framework of spatiotemporal predictive learning, in which the spatial encoder and decoder capture intra-frame features and the middle temporal module catches inter-frame correlations. While the mainstream methods employ recurrent units to capture long-term temporal dependencies, they suffer from low computational efficiency due to their unparallelizable architectures. To parallelize the temporal module, we propose the Temporal Attention Unit (TAU), which decomposes temporal attention into intra-frame statical attention and inter-frame dynamical attention. Moreover, while the mean squared error loss focuses on intra-frame errors, we introduce a novel differential divergence regularization to take inter-frame variations into account. Extensive experiments demonstrate that the proposed method enables the derived model to achieve competitive performance on various spatiotemporal prediction benchmarks.*

## 1. Introduction

The last decade has witnessed revolutionary advances in deep learning across various supervised learning tasks such as image classification [34,52,87], object detection [73,75], computational biology [21,22,43,83,84], and etc. Despite significant breakthroughs in supervised learning, which relies on large-scale labeled datasets, the potential of unsupervised learning remains largely untapped. Self-supervised learning that designs pretext tasks to produce labels derived from the data itself is recognized as a subset of unsupervised learning. In the context of self-supervised learning, *contrastive self-supervised learning* [7,8,28,33,85,86,92,104] predicts the noise contrastive estimation from predefined positive or negative pairs, and *masked self-supervised learning* [17, 32, 45, 51, 57, 102] predicts the masked patches from the visible patches. Unlike these image-level self-supervised learning, *spatiotemporal predictive learning* that predicts future frames from past frames at the video-level [6, 19, 27, 46, 60, 63].

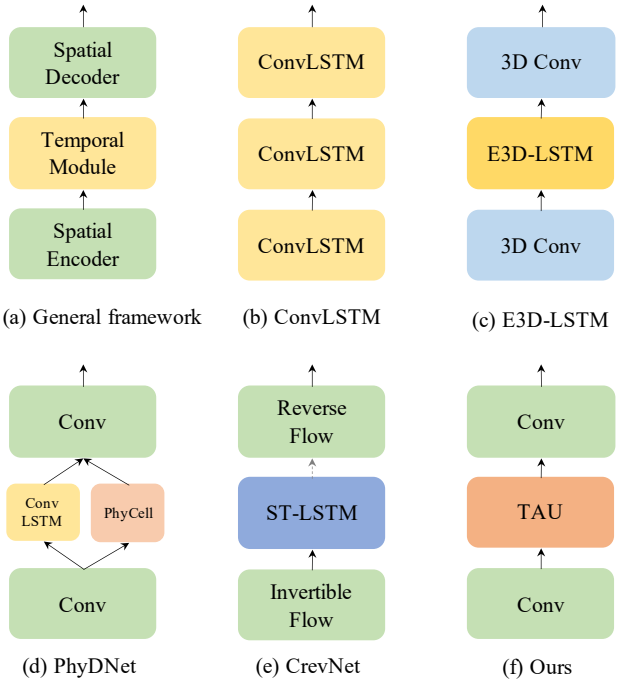


Figure 1. Comparison of model architectures among common spatiotemporal predictive learning methods. Note that we denote 2D convolutional neural networks as Conv while 3D Conv means 3D convolutional neural networks.

Accurate spatiotemporal predictive learning can benefit broad practical applications in climate change [74, 77], human motion forecasting [91, 107], traffic flow prediction [18, 97], and representation learning [39, 71]. The significance of spatiotemporal predictive learning primarily lies in its potential of exploring both spatial correlation and temporal evolution in the physical world. Moreover, the self-supervised nature of spatiotemporal predictive learning aligns well with human learning styles without a large amount of labeled data. Massive videos can provide a rich source of visual information, enabling spatiotemporal pre-

*Equal contribution.
†Corresponding author.

dictive learning to serve as a generative pre-training strategy [35, 60] for feature representation learning towards diverse downstream visual supervised tasks.

Most of the existing methods [2,3,9,11,12,19,24,29,41, 44,65,78,79,82,88,89,93–96,98,100,103] in spatiotemporal predictive learning employ hybrid architectures of convolutional neural networks and recurrent units in which spatial correlation and time evolution can be learned, respectively. Inspired by the success of long short-term memory (LSTM) [36] in sequential modeling, ConvLSTM [78] is a seminal work on the topic of spatiotemporal predictive learning that extends fully connected LSTM to convolutional LSTM towards accurate precipitation nowcasting. PredRNN [95] is an admirable work that proposes Spatiotemporal LSTM (ST-LSTM) units to model spatial appearances and temporal variations in a unified memory pool. This work provides insights on designing typical recurrent units for spatiotemporal predictive learning and inspires a series of subsequent works [4, 93, 96, 98]. E3D-LSTM [94] integrates 3D convolutional neural networks into recurrent units towards good representations with both short-term frame dependencies and long-term high-level relations. PhyDNet [29] introduces a two-branch architecture that involves physical-based PhyCells and ConvLSTMs for performing partial differential equation constraints in latent space. CrevNet [103] proposes an invertible two-way autoencoder based on flow [14, 15] and a conditionally reversible architecture for spatiotemporal predictive learning. As shown in Figure 1 (a), we present a general framework consisting of the spatial encoder/decoder and the middle temporal module, abstracted from these methods. Though these spatiotemporal predictive learning methods have different temporal modules and spatial encoders/decoders, they basically share a similar framework.

Based on the general framework, we argue that the temporal module plays an essential role in spatiotemporal predictive learning. While the common choice of the temporal module is the recurrent-based units, we explore a novel parallelizable attention module named Temporal Attention Unit (TAU) to capture time evolution. The proposed temporal attention is decomposed into intra-frame statical attention and inter-frame dynamical attention. Furthermore, we argue that the mean square error loss only focuses on intra-frame differences, and we propose a differential divergence regularization that also cares about the inter-frame variations. Keeping the spatial encoder and decoder as simple as 2D convolutional neural networks, we deliberately implement our proposed TAU modules and surprisingly find the derived model achieves competitive performance as those recurrent-based models. This observation provides a new perspective to improve spatiotemporal predictive learning by parallelizable attention networks instead of common-used recurrent units.

We conduct experiments on various datasets with different experimental settings: (1) Standard spatiotemporal predictive learning. Quantitative results on various datasets demonstrate our proposed method can achieve competitive performance on standard spatiotemporal predictive learning. (2) Generalization ability. To verify the generalization ability, we train our model on KITTI and test it on the Caltech Pedestrian dataset with different domains. (3) Predict future frames with flexible lengths. We tackle the long-length frames by feeding the predicted frames as the input and find the performance is consistently well. Through the superior performance in the above three experimental settings, we demonstrate that our proposed model can provide a novel manner that learns temporal dependencies without recurrent units.

## 2. Related works

### 2.1. Self-supervised learning

Deep learning has been well developed and applied in various fields [5, 55, 56, 109, 110]. Learning from massive data enables tremendous progress in supervised learning. By designing pretext tasks and generating labels from the data itself, self-supervised learning obtains supervisory signals. The model learns valuable representations by solving pretext tasks that leverage the underlying structure of the data. Early works on visual self-supervised learning design pretext tasks such as colorization [108], inpainting [69], rotation [26], jigsaw [66]. Contrastive self-supervised learning [7, 8, 28, 33, 92, 104] is a dominant manner in visual self-supervised learning that aims at a pretext task of grouping similar samples closer and diverse samples away from each other. However, contrastive self-supervised learning is limited by making pairs by multiple images, which affects its ability on small-scale datasets. Masked self-supervised learning [17, 32, 45, 51, 57, 102], which predicts the masked patches from the visible ones, is another research direction. Although masked pretraining has great success in natural language processing, its applications in visual tasks are challenging. Spatiotemporal predictive learning is another promising branch of self-supervised learning that focus on video-level information and predicts future frames conditioned on past frames.

In contrast to the above image-level methods, spatiotemporal predictive learning focus on video-level information and predicts future frames conditioned on past frames. By learning the intrinsic motion dynamics, the model is enabled to easily decouple the foreground and background.

### 2.2. Spatiotemporal predictive learning

Recurrent models have achieved remarkable advances in spatiotemporal predictive learning. Inspired by recurrent neural networks, VideoModeling [62] adopts language

modeling and quantizes the image patches into a large dictionary for recurrent units. ConvLSTM [78] leverages convolutional neural networks to model the LSTM architecture. PredNet [61] persistently predicts future video frames using deep recurrent convolutional neural networks with bottom-up and top-down connections. PredRNN [95] proposes a Spatiotemporal LSTM unit that extracts and memorizes spatial and temporal representations simultaneously, and its following work PredRNN++ [93] further introduces gradient highway unit and Casual LSTM to adaptively capture temporal dependencies. E3D-LSTM [94] designs eidetic memory transition in recurrent convolutional units. Conv-TT-LSTM [82] employs a higher-order ConvLSTM to predict by combining convolutional features across time. MotionRNN [100] focuses on motion trends and transient variations. LMC-Memory [50] introduces a long-term motion context memory using memory alignment learning. PredRNN-v2 [96] extends PredRNN by leveraging a memory decoupling loss and curriculum learning strategy.

Instead of using recurrent-based methods that are computationally expensive for spatiotemporal predictive learning, we introduce TAU, a model that uses visual attention mechanism to parallelize the temporal evolution without the recurrent structure. There are prior arts that have some similarities with our proposed model. PredCNN [101] and TrajectoryCNN [54] implement pure convolutional neural networks as the temporal module. SimVP [23] is a seminal work that applies blocks of Inception modules with a UNet architecture to learn the temporal evolution. Though their temporal modules are parallelizable, we argue that convolutions alone cannot capture long-term dependencies. Moreover, SimVP provides a simple baseline with minor complex attachment but a large space for further improvements. In general, SimVP first downsamples video sequences to reduce the computation, then uses Inception-UNet to learn essential spatiotemporal relationships, and upsamples the representations to predict future frames. Our work aims to replace the pivotal Inception-UNet with efficient attention modules that promote prediction performance. In this paper, we employ a simple yet effective attention mechanism to enable the temporal module not only to be parallelizable but also to capture long-term time evolution.

# 3. Methods

## 3.1. Preliminaries

We formally define the spatiotemporal predictive learning problem as follows. Given a video sequence $\mathcal{X}^{t,T} = \{\mathbf{x}^i\}_{t-T+1}^{t}$ at time $t$ with the past $T$ frames, we aim to predict the subsequent $T'$ frames $\mathcal{Y}^{t+1,T'} = \{\mathbf{x}^i\}_{t+1}^{t+1+T'}$ from time $t+1$, where $\mathbf{x}_i \in \mathbb{R}^{C\times H\times W}$ is usually an image with channels $C$, height $H$, and width $W$. In practice, we represent the video sequences as tensors, i.e., $\mathcal{X}^{t,T} \in \mathbb{R}^{T\times C\times H\times W}$ and $\mathcal{Y}^{t+1,T'} \in \mathbb{R}^{T'\times C\times H\times W}$.

The model with learnable parameters $\Theta$ learns a mapping $\mathcal{F}_\Theta : \mathcal{X}^{t,T} \mapsto \mathcal{Y}^{t+1,T'}$ by exploring both spatial and temporal dependencies. In our case, the mapping $\mathcal{F}_\Theta$ is a neural network model trained to minimize the difference between the predicted future frames and the ground-truth future frames. The optimal parameters $\Theta^*$ are:

$$\Theta^* = \arg\min_{\Theta} \mathcal{L}(\mathcal{F}_\Theta(\mathcal{X}^{t,T}), \mathcal{Y}^{t+1,T'}), \tag{1}$$

where $\mathcal{L}$ is a loss function that evaluates such differences.

## 3.2. Overview

We illustrate the overview model in Figure 2 using the input Moving MNIST [81] data as an example.

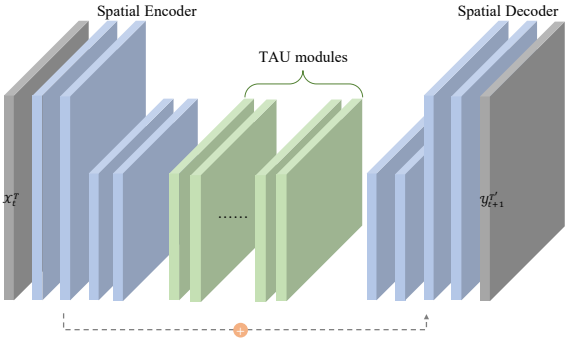


Figure 2. The overview architecture of our proposed model.

Striving for simplicity, the model follows the general framework in Figure 1, while the spatial encoder consists of four vanilla 2D convolutional layers, and the spatial decoder consists of four 2D transposed convolutional layers ('Conv2d' and 'ConvTranspose2d' in PyTorch respectively). We add a residual connection from the first convolutional layer to the last transposed convolutional layer for preserving the spatial-dependent features. Stacks of TAU modules are in the middle of the spatial encoder and decoder to extract temporal-dependent features. Though our model is simple, it can efficiently learn both spatial-dependent and temporal-dependent features without recurrent architectures.

## 3.3. Temporal Attention Unit

Suppose a batch of input video tensors $\mathcal{B} \in \mathbb{R}^{B\times T\times C\times H\times W}$ with the number of videos $B = |\mathcal{B}|$ is given. In the spatial encoder and decoder, we reshape the sequential input data $B \times T \times C \times H \times W$ as $(B \times T) \times C \times H \times W$ so that only spatial correlations are taken into account. In the temporal module, we reshape the feature $B \times T \times C \times H \times W$ as $B \times (T \times C) \times H \times W$ so that frames are arranged in order on the channel dimension.

We decompose the temporal attention into the intra-frame statical attention and the inter-frame dynamical attention, as shown in Figure 3. Inspired by the recent progress of vision Transformers (ViTs) [17, 57] and large kernel convolutions [13, 30, 58], we propose to employ small kernel depth-wise convolutions (DW Conv), depth-wise convolutions with dilations (DW-D Conv), and $1\times1$ convolutions to model the large kernel convolutions. Through the obtained large receptive field on intra-frames, the statical attention is able to capture long-range dependencies. However, the statical attention alone is not enough for learning temporal evolutions along the timeline. Thus, we employ the dynamical attention that learns the attention weights of channels in a squeeze-and-excitation manner [38]. The final attention is the product of dynamical attention and statical attention.

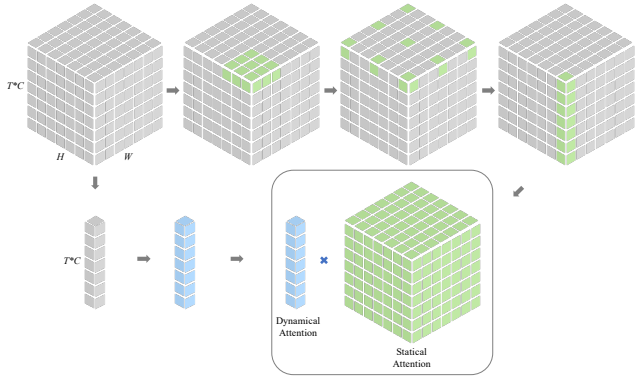


Figure 3. The intra-frame statical attention and the inter-frame dynamical attention.

We show the detailed scheme of our model in Figure 4. The proposed TAU module can be formally expressed as:

$$\begin{aligned} \text{SA} &= \text{Conv}_{1\times1}(\text{DW-D Conv}(\text{DW Conv}(H))), \\ \text{DA} &= \text{FC}(\text{AvgPool}(H)), \\ H' &= (\text{SA} \otimes \text{DA}) \odot H, \end{aligned} \tag{2}$$

where $H \in \mathbb{R}^{B\times(T\times C')\times H\times W}$ is the hidden feature that will be fed into the TAU module, $\text{SA} \in \mathbb{R}^{B\times(T\times C')\times H\times W}$, $\text{DA} \in \mathbb{R}^{B\times(T\times C')\times1\times1}$ denote the statical and dynamical attention, FC and AvgPool are fully connected layers and the average pooling. We represent the Kronecker product by $\otimes$ and the Hadamard product by $\odot$.

### 3.4. Differential Divergence Regularization

To improve the prediction of our model, we further propose a differential divergence regularization that forces the model to learn the differences between consecutive frames and be aware of the inherent variation.

Given the predicted frames $\widehat{\mathcal{Y}} = \mathcal{F}_\Theta(\mathcal{X}) \in \mathbb{R}^{T'\times C\times H\times W}$ and its corresponding ground-truth frames

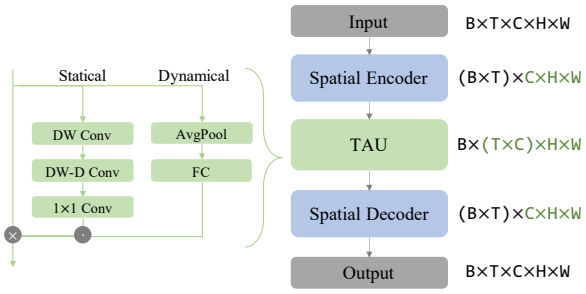


Figure 4. The detailed schema of our model.

$\mathcal{Y}$, we first calculate their forward difference $\Delta\widehat{\mathcal{Y}}, \Delta\mathcal{Y} \in \mathbb{R}^{(T'-1)\times C\times H\times W}$, where:

$$\begin{aligned} \Delta\widehat{\mathcal{Y}}_i &= \widehat{\mathcal{Y}}_{i+1} - \widehat{\mathcal{Y}}_i, \\ \Delta\mathcal{Y}_i &= \mathcal{Y}_{i+1} - \mathcal{Y}_i. \end{aligned} \tag{3}$$

Then, we transform the differences into probabilities by the softmax function on the channel, height, and width dimension and obtain $\sigma(\Delta\widehat{\mathcal{Y}}), \sigma(\Delta\mathcal{Y})$, where:

$$\begin{aligned} \sigma(\Delta\widehat{\mathcal{Y}})_{i,j,k,l} &= \frac{\exp\left(\Delta\widehat{\mathcal{Y}}_{i,j,k,l}/\tau\right)}{\sum_{j'=1}^{C}\sum_{k'=1}^{H}\sum_{l'=1}^{W}\exp(\Delta\widehat{\mathcal{Y}}_{i,j',k',l'}/\tau)}, \\ \sigma(\Delta\mathcal{Y})_{i,j,k,l} &= \frac{\exp\left(\Delta\mathcal{Y}_{i,j,k,l}\right)/\tau}{\sum_{j'=1}^{C}\sum_{k'=1}^{H}\sum_{l'=1}^{W}\exp(\Delta\mathcal{Y}_{i,j',k',l'}/\tau)}, \end{aligned} \tag{4}$$

and $\tau$ represents the temperature parameter which we empirically set as 0.1 to sharpen the probability distribution. Through the competition mechanism of the softmax function [10, 70, 99], the high difference frames are penalized.

Thus, the differential divergence regularization $\mathcal{L}_{reg}$ is defined as the Kullback-Leibler divergence between the probability distributions $\sigma(\Delta\widehat{\mathcal{Y}})$ and $\sigma(\Delta\mathcal{Y})$:

$$\begin{aligned} \mathcal{L}_{reg}(\widehat{\mathcal{Y}}, \mathcal{Y}) &= D_{KL}(\sigma(\Delta\widehat{\mathcal{Y}}) \,||\, \sigma(\Delta\mathcal{Y})) \\ &= \sum_{i=1}^{T'-1} \sigma(\Delta\widehat{\mathcal{Y}}_i) \log \frac{\sigma(\Delta\widehat{\mathcal{Y}}_i)}{\sigma(\Delta\mathcal{Y}_i)}. \end{aligned} \tag{5}$$

Our model is trained end-to-end in a fully unsupervised manner, and the overall objective function consists of the mean square error loss and the differential divergence regularization weighted by a constant $\alpha$:

$$\mathcal{L} = \sum_{i=1}^{T'} \|\widehat{\mathcal{Y}} - \mathcal{Y}\|^2 + \alpha\mathcal{L}_{reg}(\widehat{\mathcal{Y}}, \mathcal{Y}), \tag{6}$$

where the first term focuses on intra-frame-level differences, and the second regularization term focuses on inter-frame-level variations.

# 4. Experiments

In this section, we present experiments that demonstrate the effectiveness of our proposed method. The experiments are conducted on various datasets with different settings to validate our proposed model from three aspects:

- Standard spatiotemporal predictive learning (Section 4.2). We recognize the prediction problem of the same number of input and output frames as the standard spatiotemporal predictive learning. We evaluate the performance on standard spatiotemporal predictive learning and compare our model with state-of-the-art methods with Moving MNIST [81]and TaxiBJ [106] datasets.
- Generalization ability across different datasets (Section 4.3). Generalizing the learned knowledge to other domains is a challenge in unsupervised learning. We investigate the such ability of our method by training the model on the KITTI [25] dataset and evaluating it on the Caltech Pedestrian [16] dataset.
- Predicting frames with flexible lengths (Section 4.4). One of the advantages of recurrent units is that they can easily handle flexible-length frames like the KTH dataset [76]. Our work tackles the long-length frame prediction by imitating recurrent units that feed predicted frames as the input and recursively produce long-term predictions.

## 4.1. Experimental Setups

**Datasets** We quantitatively evaluate our model on the following datasets for both synthetic and real-world scenarios:

- **Moving MNIST** [81] is a synthetic dataset consisting of two digits independently moving within the 64 $\times$ 64 grid and bouncing off the boundary. It is a standard benchmark in spatiotemporal predictive learning.
- **TaxiBJ** contains the trajectory data in Beijing collected from taxicab GPS with two channels, i.e., inflow or outflow defined in [106]. Following the previous works [29, 98], we normalize the data into $[0, 1]$.
- **KTH** [76] contains 25 individuals performing six types of actions. Following [90, 94], we use person 1-16 for training and 17-25 for testing. Models are trained to predict the next 20 or 40 frames from the previous 10 observations.
- **Caltech Pedestrian** is a driving dataset focusing on detecting pedestrians. It consists of approximately 10 hours of $640 \times 480$ videos taken from vehicles driving through regular traffic in an urban environment. We follow the same protocol of PredNet [61] and CrevNet [103] for pre-processing, training, and evaluation.

We summarize the statistics of the above datasets in Table 1, including the number of training samples $N_{train}$ and the number of testing samples $N_{test}$.

Table 1. The statistics of datasets. The training or testing set has $N_{train}$ or $N_{test}$ samples, composed by $T$ or $T'$ images with the shape $(C, H, W)$.

| | $N_{train}$ | $N_{test}$ | $(C, H, W)$ | $T$ | $T'$ |
|---|---|---|---|---|---|
| MMNIST | 10000 | 10000 | (1, 64, 64) | 10 | 10 |
| TaxiBJ | 19627 | 1334 | (2, 32, 32) | 4 | 4 |
| KTH | 5200 | 3167 | (1, 128, 128) | 10 | 20 or 40 |
| Caltech | 2042 | 1983 | (3, 128, 160) | 10 | 1 |

**Measurement** Following [29, 103], we employ Mean Squared Error (MSE), Mean Absolute Error (MAE), Structure Similarity Index Measure (SSIM), and Peak Signal to Noise Ratio (PSNR) to evaluate the quality of predictions. MSE and MAE estimate the absolute pixel-wise errors, SSIM measures the similarity of structural information within the spatial neighborhoods, and PSNR is an expression for the ratio between the maximum possible power of a signal and the power of distorted noise.

**Implementation details** We implement the proposed method with the Pytorch framework and conduct experiments on a single NVIDIA-V100 GPU. The model is trained with a mini-batch of 16 video sequences while the AdamW optimizer is utilized with a learning rate of 0.01 and a weight decay of 0.05.

## 4.2. Standard spatiotemporal predictive learning

### 4.2.1 Moving MNIST

This dataset is a standard benchmark in spatiotemporal predictive learning. We evaluate our proposed method against strong recent baselines, including competitive recurrent architectures: ConvLSTM [78], PredRNN [95], PredRNN++ [93], MIM [98], LMC [50], E3D-LSTM [94], Conv-TT-LSTM [82], and CrevNet [103]. We also compare the method DDPAE [37] that is specifically designed for this dataset. The quantitative results are reported in Table 2, and qualitative visualizations of the predicted results are shown in Figure 5.

Our proposed method significantly outperforms all the baselines above under three different metrics. The performance gain is large with respect to state-of-the-art recurrent methods.

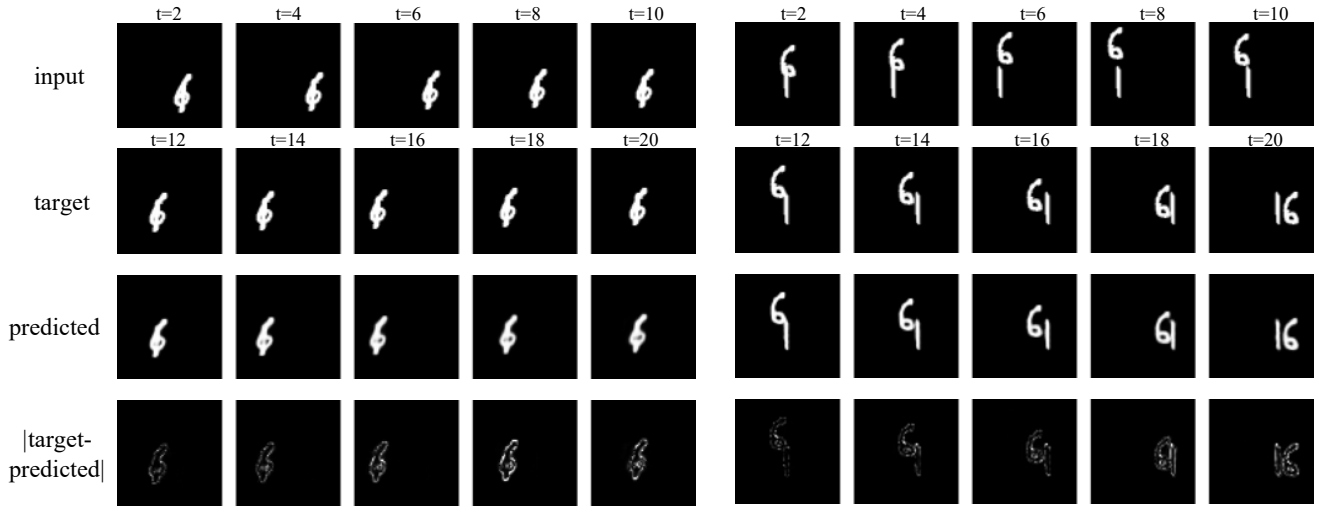


Figure 5. Qualitative visualization of predicted results on Moving MNIST dataset. The differences between the ground truth and the predicted frames are visualized in the last row.

Table 2. Quantitative results of different methods on the Moving MNIST dataset (10 → 10 frames).

| Method | Moving MNIST MSE↓ | MAE↓ | SSIM↑ |
|---|---|---|---|
| ConvLSTM [78] | 103.3 | 182.9 | 0.707 |
| VPN [44] | 64.1 | - | 0.870 |
| PredRNN [95] | 56.8 | 126.1 | 0.867 |
| PredRNN++ [93] | 46.5 | 106.8 | 0.898 |
| MIM [98] | 44.2 | 101.1 | 0.910 |
| LMC [50] | 41.5 | - | 0.924 |
| E3D-LSTM [94] | 41.3 | 87.2 | 0.910 |
| Conv-TT-LSTM [82] | 53.0 | - | 0.915 |
| DDPAE [37] | 38.9 | 90.7 | 0.922 |
| PhyDNet [29] | 24.4 | 70.3 | 0.947 |
| SimVP [23] | 23.8 | 68.9 | 0.948 |
| Crevnet [103] | 22.3 | - | 0.949 |
| Ours | **19.8** | **60.3** | **0.957** |

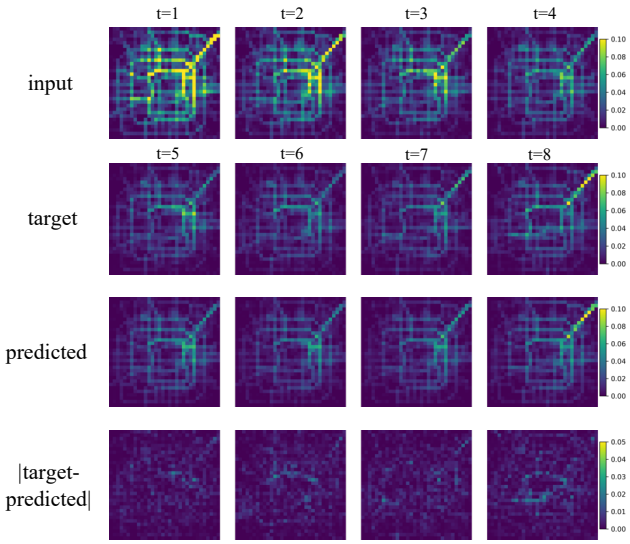


Figure 6. Qualitative visualization of predicted results on TaxiBJ dataset. The differences between the ground truth and the predicted frames are visualized in the last row.

#### 4.2.2 TaxiBJ

We evaluate our proposed model on a complicated real-world dataset, TaxiBJ [106]. Driven by human consciousness, the complex real-world traffic flows requires modeling transport phenomena and traffic diffusion for prediction. Due to the spatiotemporal nature of the traffic forecasting task, we straightforwardly implement our model for it.

The qualitative visualizations of the predicted results are shown in Figure 6, and the quantitative results are reported in Table 3. Though the given frames are quite different from the future frames, our model can still accurately produce reliable frames. The difference between target frames and predicted frames is mainly located in central spots, but the overall trend is approximating the ground truth. It can also be observed that the quantitative results are consistently well, which demonstrates the practical application value in real-world scenarios.

### 4.3. Generalization across different datasets

The generalization ability is one of the fundamental problems in artificial intelligence technology. Traditional supervised learning suffers from its poor generalization of

Table 3. Quantitative results of different methods on the TaxiBJ dataset (4 → 4 frames).

| Method | TaxiBJ MSE × 100↓ | MAE↓ | SSIM↑ |
|---|---|---|---|
| ConvLSTM [78] | 48.5 | 17.7 | 0.978 |
| PredRNN [95] | 46.4 | 17.1 | 0.971 |
| PredRNN++ [93] | 44.8 | 16.9 | 0.977 |
| MIM [98] | 42.9 | 16.6 | 0.971 |
| E3D-LSTM [94] | 43.2 | 16.9 | 0.979 |
| PhyDNet [29] | 41.9 | 16.2 | 0.982 |
| SimVP [23] | 41.4 | 16.2 | 0.982 |
| Ours | **34.4** | **15.6** | **0.983** |

labeled datasets with different domains. Self-supervised learning aims to learn robust representations based on unlabeled data and evaluates the generalization ability of the learned model. While contrastive self-supervised learning and masked self-supervised learning in visual tasks usually evaluate such generalization ability by downstream tasks, we evaluate it by the prediction results across different datasets in spatiotemporal predictive learning.


Figure 7. Qualitative visualization of predicted results on Caltech dataset. The differences between the ground truth and the predicted frames are visualized in the last row.

Following the previous works [53, 61, 103], we train the model using the raw video sequences from the KITTI dataset and evaluate the model by Caltech Pedestrian dataset that is made to match the frame rate and image size (128 × 160) of the KTTI dataset. The final prediction is made for the next frame after a 10-frame warm-up.

We show the qualitative visualization in Figure 7 and report the quantitative results in Table 4. Note that some of the baseline results are copied from [68]. It can be seen that our proposed method achieves state-of-the-art performance under both SSIM and PSNR metrics in this generalization evaluation task. Moreover, our model shows significantly robust predictions in terms of the variation of illumination and lane line, suggesting its practical applications in autonomous vehicles.

Table 4. Quantitative results of different methods on the Caltech Pedestrian dataset (10 → 1 frame).

| Method | Caltech Pedestrian SSIM↑ | PSNR↑ |
|---|---|---|
| BeyondMSE [64] | 0.847 | - |
| MCnet [90] | 0.879 | - |
| DVF [59] | 0.897 | 26.2 |
| Dual-GAN [53] | 0.899 | - |
| CtrlGen [31] | 0.900 | 26.5 |
| PredNet [61] | 0.905 | 27.6 |
| ContextVP [4] | 0.921 | 28.7 |
| GAN-VGG [80] | 0.916 | - |
| G-VGG [80] | 0.917 | - |
| SDC-Net [72] | 0.918 | - |
| rCycleGan [47] | 0.919 | 29.2 |
| DPG [20] | 0.923 | 28.2 |
| G-MAE [80] | 0.923 | - |
| GAN-MAE [80] | 0.923 | - |
| CrevNet [103] | 0.925 | 29.3 |
| STMFANet [41] | 0.927 | 29.1 |
| SimVP [23] | 0.940 | 33.1 |
| Ours | **0.946** | **33.7** |

### 4.4. Predicting frames with flexible lengths

Though recurrent units are adept at handling flexible-length frames, our model can also easily tackle such problems by imitating recurrent units that feed predicted frames as the input and recursively produce predictions. For this KTH dataset, our model is trained to predict the next 20 or 40 frames from the given 10 observations. Moreover, this dataset contains six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios: outdoors, outdoors with scale variations, outdoors with different clothes and indoors. The difficulty of this human motion prediction task not only lies in its flexible lengths of predicted frames but also in its complex dynamics involving the randomness of human consciousness. Following [95], we use the Peak Signal Noise Ratio (PSNR) and the Structural Similarity Index Measure (SSIM) as evaluation metrics to measure the framewise prediction quality from the perceptive view. The detailed quantitative results are shown in Table 5.

Table 5. Quantitative results of different methods on the KTH dataset ($10 \rightarrow 20$ and $10 \rightarrow 40$ frames).

| Method | KTH ($10 \rightarrow 20$) | | KTH ($10 \rightarrow 40$) | |
|---|---|---|---|---|
| | SSIM↑ | PSNR↑ | SSIM↑ | PSNR↑ |
| MCnet [90] | 0.804 | 25.95 | 0.73 | 23.89 |
| ConvLSTM [78] | 0.712 | 23.58 | 0.639 | 22.85 |
| SAVP [48] | 0.746 | 25.38 | 0.701 | 23.97 |
| VPN [44] | 0.746 | 23.76 | – | – |
| DFN [40] | 0.794 | 27.26 | 0.652 | 23.01 |
| fRNN [67] | 0.771 | 26.12 | 0.678 | 23.77 |
| Znet [105] | 0.817 | 27.58 | – | – |
| SV2Pv [1] | 0.838 | 27.79 | 0.789 | 26.12 |
| PredRNN [95] | 0.839 | 27.55 | 0.703 | 24.16 |
| VarNet [42] | 0.843 | 28.48 | 0.739 | 25.37 |
| SAVP-VAE [48] | 0.852 | 27.77 | 0.811 | 26.18 |
| PredRNN++ [93] | 0.865 | 28.47 | 0.741 | 25.21 |
| MSNET [49] | 0.876 | 27.08 | – | – |
| E3d-LSTM [94] | 0.879 | 29.31 | 0.810 | 27.24 |
| STMFANet [41] | 0.893 | 29.85 | 0.851 | 27.56 |
| SimVP [23] | 0.905 | 33.72 | 0.886 | 32.93 |
| Ours | **0.911** | **34.13** | **0.897** | **33.01** |

### 4.5. Empirical Running Time

TAU benefits from the parallelizable computation architecture, which leads to fast convergence and high training speed. We evaluate our efficiency by measuring the running time against state-of-the-art spatiotemporal predictive learning methods.

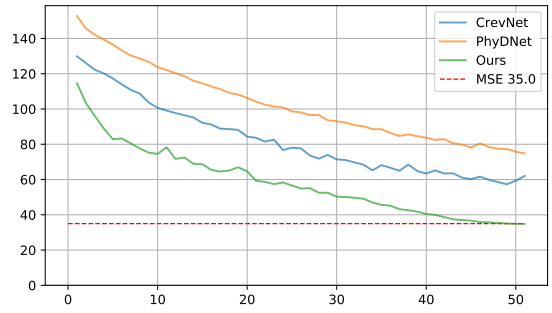


Figure 8. The learning curve comparison between state-of-the-art methods and ours (evaluated by MSE). The red dotted line denotes MSE 35.0, and only the first 50 epochs are shown.

The experiments are conducted on a single Tesla V100 GPU, and the batch size is set as 16. For the Moving MNIST dataset, CrevNet [103] needs about 30 minutes per epoch, and PhyDNet [29] needs about 7 minutes per epoch. Our model only requires 2.5 minutes per epoch. Furthermore, our method is able to converge at a rapid rate. As shown in Figure 8, on the Moving MNIST dataset, our model can achieve MSE 35.0 with only 50 epochs, while CrevNet and PhyDNet are far from this performance.

### 4.6. Computational Cost and Ablation Study

We compare the performance and computational cost with state-of-the-art methods in the first several rows in Table 6. Our model achieves superior results with better performance and much lower Flops. We also conduct ablation studies and summarize the results in Table 6. It can be seen that replacing TAU with the same number of convolutional blocks with vanilla $3 \times 3$ convolutions (Conv Baseline) significantly degrades the performance. Training our model without differential divergence regularization (w/o DDR) will also weaken the prediction results. Both the TAU module and differential divergence regularization are useful. We also find that SA and DA of the TAU module play important roles in better performance.

Table 6. Ablation study of our proposed method.

| Method | M-MNIST | TaxiBJ ($\times 100$) | Flops(G) |
|---|---|---|---|
| PredRNN | 56.8 | 46.4 | 115.94 |
| PredRNN++ | 46.4 | 44.8 | 171.73 |
| MIM | 44.2 | 42.9 | 179.17 |
| E3DLSTM | 41.3 | 43.2 | 298.85 |
| SimVP | 23.8 | 41.4 | 19.43 |
| Conv Baseline | 58.9 | 43.5 | 6.11 |
| Ours w/o SA | 23.2 | 40.8 | 15.30 |
| Ours w/o DA | 22.4 | 38.4 | 15.96 |
| Ours w/o DDR | 21.1 | 37.7 | 15.96 |
| Ours | **19.8** | **34.4** | 15.96 |

## 5. Conclusion

In this paper, we present a general framework of spatiotemporal predictive learning and propose an attention-based temporal module to replace the common-used recurrent units. By decomposing the temporal module into the intra-frame statical attention and the inter-frame dynamical attention, our proposed TAU module can achieve competitive performance across various experimental settings and datasets. Moreover, a novel differential divergence regularization is proposed to overcome the drawback of MSE loss that only considers the intra-frame error. In summary, our work highlights the importance of both intra-frame and inter-frame variations that enable the model to capture long-term relations and provide a new paradigm of efficient spatiotemporal predictive learning.

**Acknowledgement.** We thank the anonymous reviewers for their constructive and helpful reviews. This work was supported by the National Key R&D Program of China (2022ZD0115100), the National Natural Science Foundation of China (U21A20427), the Competitive Research Fund (WU2022A009) from the Westlake Center for Synthetic Biology and Integrated Bioengineering

## References


[1] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018. 8

[2] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and Dumitru Erhan. Fitvid: Overfitting in pixel-level video prediction. *arXiv preprint arXiv:2106.13195*, 2021. 2

[3] Sarthak Bhagat, Shagun Uppal, Zhuyun Yin, and Nengli Lim. Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. In *European Conference on Computer Vision*, pages 102–117. Springer, 2020. 2

[4] Wonmin Byeon, Qin Wang, Rupesh Kumar Srivastava, and Petros Koumoutsakos. Contextvp: Fully context-aware video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 753–769, 2018. 2, 7

[5] Hanqun Cao, Cheng Tan, Zhangyang Gao, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646*, 2022. 2

[6] Lluis Castrejon, Nicolas Ballas, and Aaron Courville. Improved conditional vrnns for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7608–7617, 2019. 1

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020. 1, 2

[8] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33:22243–22255, 2020. 1, 2

[9] Yu Cheng, Bo Wang, Bo Yang, and Robby T Tan. Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 4, page 12, 2021. 2

[10] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2020. 4

[11] Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *International Conference on Machine Learning*, pages 1174–1183. PMLR, 2018. 2

[12] Emily L Denton et al. Unsupervised learning of disentangled representations from video. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[13] Xiaohan Ding, Xiangyu Zhang, Yizhuang Zhou, Jungong Han, Guiguang Ding, and Jian Sun. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. *arXiv preprint arXiv:2203.06717*, 2022. 4

[14] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. In *International Conference on Learning Representations Workshop*, 2015. 2

[15] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. In *International Conference on Learning Representations*, 2017. 2

[16] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, pages 304–311. IEEE, 2009. 5

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020. 1, 2, 4

[18] Shen Fang, Qi Zhang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Gstnet: Global spatial-temporal network for traffic flow prediction. In *IJCAI*, pages 2286–2293, 2019. 1

[19] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in Neural Information Processing Systems*, 29, 2016. 1, 2

[20] Hang Gao, Huazhe Xu, Qi-Zhi Cai, Ruth Wang, Fisher Yu, and Trevor Darrell. Disentangling propagation and generation for video prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9006–9015, 2019. 7

[21] Zhangyang Gao, Cheng Tan, Stan Li, et al. Alphadesign: A graph protein design method and benchmark on alphafolddb. *arXiv preprint arXiv:2202.01079*, 2022. 1

[22] Zhangyang Gao, Cheng Tan, and Stan Z Li. Pifold: Toward effective and efficient protein inverse folding. *arXiv preprint arXiv:2209.12643*, 2022. 1

[23] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3170–3180, June 2022. 3, 6, 7, 8

[24] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR, 2017. 2

[25] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 5

[26] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 2

[27] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative varia-

tional inference. In *International Conference on Machine Learning*, pages 2424–2433. PMLR, 2019. 1

[28] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21271–21284. Curran Associates, Inc., 2020. 1, 2

[29] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020. 2, 5, 6, 7, 8

[30] Meng-Hao Guo, Cheng-Ze Lu, Zheng-Ning Liu, Ming-Ming Cheng, and Shi-Min Hu. Visual attention network. *arXiv preprint arXiv:2202.09741*, 2022. 4

[31] Zekun Hao, Xun Huang, and Serge Belongie. Controllable video generation with sparse trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7854–7863, 2018. 7

[32] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[33] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020. 1, 2

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[35] Geoffrey Hinton. How to represent part-whole hierarchies in a neural network. *arXiv preprint arXiv:2102.12627*, 2021. 2

[36] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[37] Jun-Ting Hsieh, Bingbin Liu, De-An Huang, Li F Fei-Fei, and Juan Carlos Niebles. Learning to decompose and disentangle representations for video prediction. *Advances in Neural Information Processing Systems*, 31, 2018. 5, 6

[38] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 4

[39] Simon Jenni, Givi Meishvili, and Paolo Favaro. Video representation learning by recognizing temporal transformations. In *European Conference on Computer Vision*, pages 425–442. Springer, 2020. 1

[40] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016. 8

[41] Beibei Jin, Yu Hu, Qiankun Tang, Jingyu Niu, Zhiping Shi, Yinhe Han, and Xiaowei Li. Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4554–4563, 2020. 2, 7, 8

[42] Beibei Jin, Yu Hu, Yiming Zeng, Qiankun Tang, Shice Liu, and Jing Ye. Varnet: Exploring variations for unsupervised video prediction. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5801–5806. IEEE, 2018. 8

[43] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. 1

[44] Nal Kalchbrenner, Aäron Oord, Karen Simonyan, Ivo Danihelka, Oriol Vinyals, Alex Graves, and Koray Kavukcuoglu. Video pixel networks. In *International Conference on Machine Learning*, pages 1771–1779. PMLR, 2017. 2, 6, 8

[45] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. 1, 2

[46] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. 1

[47] Yong-Hoon Kwon and Min-Gyu Park. Predicting future frames using retrospective cycle gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1811–1820, 2019. 7

[48] Alex X Lee, Richard Zhang, Frederik Ebert, Pieter Abbeel, Chelsea Finn, and Sergey Levine. Stochastic adversarial video prediction. *arXiv preprint arXiv:1804.01523*, 2018. 8

[49] Jungbeom Lee, Jangho Lee, Sungmin Lee, and Sungroh Yoon. Mutual suppression network for video prediction using disentangled features. *arXiv preprint arXiv:1804.04810*, 2018. 8

[50] Sangmin Lee, Hak Gu Kim, Dae Hwi Choi, Hyung-Il Kim, and Yong Man Ro. Video prediction recalling long-term motion context via memory alignment learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3054–3063, 2021. 3, 5, 6

[51] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020. 1, 2

[52] Siyuan Li, Zedong Wang, Zicheng Liu, Cheng Tan, Haitao Lin, Di Wu, Zhiyuan Chen, Jiangbin Zheng, and Stan Z Li. Efficient multi-order gated aggregation network. *arXiv preprint arXiv:2211.03295*, 2022. 1

[53] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *proceedings of the IEEE international conference on computer vision*, pages 1744–1752, 2017. 7

[54] Xiaoli Liu, Jianqin Yin, Jin Liu, Pengxiang Ding, Jun Liu, and Huaping Liu. Trajectorycnn: a new spatio-temporal feature learning network for human motion prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(6):2133–2146, 2020. 3

[55] Zicheng Liu, Siyuan Li, Ge Wang, Cheng Tan, Lirong Wu, and Stan Z Li. Decoupled mixup for data-efficient learning. *arXiv preprint arXiv:2203.10761*, 2022. 2

[56] Zicheng Liu, Siyuan Li, Di Wu, Zihan Liu, Zhiyuan Chen, Lirong Wu, and Stan Z Li. Automix: Unveiling the power of mixup for stronger classifiers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 441–458. Springer, 2022. 2

[57] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1, 2, 4

[58] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *arXiv preprint arXiv:2201.03545*, 2022. 4

[59] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017. 7

[60] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019. 1, 2

[61] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. In *International Conference on Learning Representations*, 2017. 3, 5, 7

[62] Arthur Szlam Marc'Aurelio Ranzato, Joan Bruna, Michaël Mathieu, Ronan Collobert, and Sumit Chopra. Video (language) modeling: a baseline for generative models of natural videos. *CoRR, abs/1412.6604*, 2, 2014. 2

[63] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019. 1

[64] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015. 7

[65] Matthias Minderer, Chen Sun, Ruben Villegas, Forrester Cole, Kevin P Murphy, and Honglak Lee. Unsupervised learning of object structure and dynamics from videos. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[66] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[67] Marc Oliu, Javier Selva, and Sergio Escalera. Folded recurrent neural networks for future video prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 716–731, 2018. 8

[68] Sergiu Oprea, Pablo Martinez-Gonzalez, Alberto Garcia-Garcia, John Alejandro Castro-Vargas, Sergio Orts-Escolano, Jose Garcia-Rodriguez, and Antonis Argyros. A review on deep learning techniques for video prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 7

[69] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[70] Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah Smith, and Lingpeng Kong. Random feature attention. In *International Conference on Learning Representations*, 2020. 4

[71] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2021. 1

[72] Fitsum A Reda, Guilin Liu, Kevin J Shih, Robert Kirby, Jon Barker, David Tarjan, Andrew Tao, and Bryan Catanzaro. Sdc-net: Video prediction using spatially-displaced convolution. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 718–733, 2018. 7

[73] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 1

[74] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, et al. Deep learning and process understanding for data-driven earth system science. *Nature*, 566(7743):195–204, 2019. 1

[75] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28:91–99, 2015. 1

[76] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004. 5

[77] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28, 2015. 1

[78] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm

network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28, 2015. 2, 3, 5, 6, 7, 8

[79] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in Neural Information Processing Systems*, 30, 2017. 2

[80] Osamu Shouno. Photo-realistic video prediction on natural videos of largely changing frames. *arXiv preprint arXiv:2003.08635*, 2020. 7

[81] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852. PMLR, 2015. 3, 5

[82] Jiahao Su, Wonmin Byeon, Jean Kossaifi, Furong Huang, Jan Kautz, and Anima Anandkumar. Convolutional tensor-train lstm for spatio-temporal learning. *Advances in Neural Information Processing Systems*, 33:13714–13726, 2020. 2, 3, 5, 6

[83] Cheng Tan, Zhangyang Gao, and Stan Z Li. Rfold: Towards simple yet effective rna secondary structure prediction. *arXiv preprint arXiv:2212.14041*, 2022. 1

[84] Cheng Tan, Zhangyang Gao, and Stan Z Li. Target-aware molecular graph generation. *arXiv preprint arXiv:2202.04829*, 2022. 1

[85] Cheng Tan, Zhangyang Gao, Lirong Wu, Siyuan Li, and Stan Z Li. Hyperspherical consistency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7244–7255, 2022. 1

[86] Cheng Tan, Jun Xia, Lirong Wu, and Stan Z Li. Co-learning: Learning from noisy labels with self-supervision. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1405–1413, 2021. 1

[87] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019. 1

[88] Ruben Villegas, Dumitru Erhan, Honglak Lee, et al. Hierarchical long-term video prediction without supervision. In *International Conference on Machine Learning*, pages 6038–6046. PMLR, 2018. 2

[89] Ruben Villegas, Arkanath Pathak, Harini Kannan, Dumitru Erhan, Quoc V Le, and Honglak Lee. High fidelity video prediction with large stochastic recurrent neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[90] Ruben Villegas, Jimei Yang, Seunghoon Hong, Xunyu Lin, and Honglak Lee. Decomposing motion and content for natural video sequence prediction. In *International Conference on Learning Representations*, 2017. 5, 7, 8

[91] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. Rgb-d-based human motion recognition with deep learning: A survey. *Computer Vision and Image Understanding*, 171:118–139, 2018. 1

[92] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 1, 2

[93] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and S Yu Philip. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *International Conference on Machine Learning*, pages 5123–5132. PMLR, 2018. 2, 3, 5, 6, 7, 8

[94] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *International conference on learning representations*, 2018. 2, 3, 5, 6, 7, 8

[95] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in Neural Information Processing Systems*, 30, 2017. 2, 3, 5, 6, 7, 8

[96] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *arXiv preprint arXiv:2103.09504*, 2021. 2, 3

[97] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip 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*, pages 9154–9162, 2019. 1

[98] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip 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*, pages 9154–9162, 2019. 2, 5, 6, 7

[99] Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 24226–24242. PMLR, 17–23 Jul 2022. 4

[100] Haixu Wu, Zhiyu Yao, Jianmin Wang, and Mingsheng Long. Motionrnn: A flexible model for video prediction with spacetime-varying motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15435–15444, 2021. 2, 3

[101] Ziru Xu, Yunbo Wang, Mingsheng Long, Jianmin Wang, and M KLiss. Predcnn: Predictive learning with cascade convolutions. In *IJCAI*, pages 2940–2947, 2018. 3

[102] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[103] Wei Yu, Yichao Lu, Steve Easterbrook, and Sanja Fidler. Efficient and information-preserving future frame prediction and beyond. In *International Conference on Learning Representations*, 2019. 2, 5, 6, 7, 8

[104] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via

redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 1, 2

[105] Jianjin Zhang, Yunbo Wang, Mingsheng Long, Wang Jianmin, and S Yu Philip. Z-order recurrent neural networks for video prediction. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 230–235. IEEE, 2019. 8

[106] Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 5, 6

[107] Liang Zhang, Guangming Zhu, Peiyi Shen, Juan Song, Syed Afaq Shah, and Mohammed Bennamoun. Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3120–3128, 2017. 1

[108] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 2

[109] Jiangbin Zheng, Yidong Chen, Chong Wu, Xiaodong Shi, and Suhail Muhammad Kamal. Enhancing neural sign language translation by highlighting the facial expression information. *Neurocomputing*, 464:462–472, 2021. 2

[110] Jiangbin Zheng, Yile Wang, Ge Wang, Jun Xia, Yufei Huang, Guojiang Zhao, Yue Zhang, and Stan Li. Using context-to-vector with graph retrofitting to improve word embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8154–8163, 2022. 2