Triplet Attention Transformer for Spatiotemporal Predictive Learning

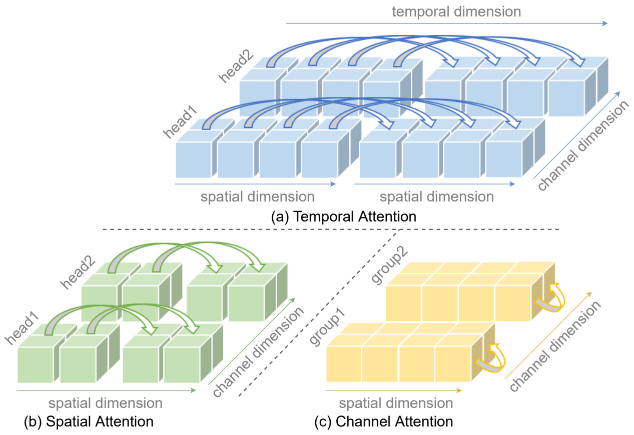
Xuesong Nie[‡] Xi Chen[†] Haoyuan Jin[‡] Zhihang Zhu[‡] Yunfeng Yan[‡◇] Donglian Qi[‡]
[‡]Zhejiang University
[†]The University of Hong Kong

Abstract

Spatiotemporal predictive learning offers a self-supervised learning paradigm that enables models to learn both spatial and temporal patterns by predicting future sequences based on historical sequences. Mainstream methods are dominated by recurrent units, yet they are limited by their lack of parallelization and often underperform in real-world scenarios. To improve prediction quality while maintaining computational efficiency, we propose an innovative triplet attention transformer designed to capture both inter-frame dynamics and intra-frame static features. Specifically, the model incorporates the Triplet Attention Module (TAM), which replaces traditional recurrent units by exploring self-attention mechanisms in temporal, spatial, and channel dimensions. In this configuration: (i) temporal tokens contain abstract representations of inter-frame, facilitating the capture of inherent temporal dependencies; (ii) spatial and channel attention combine to refine the intra-frame representation by performing fine-grained interactions across spatial and channel dimensions. Alternating temporal, spatial, and channel-level attention allows our approach to learn more complex short- and long-range spatiotemporal dependencies. Extensive experiments demonstrate performance surpassing existing recurrent-based and recurrent-free methods, achieving state-of-the-art under multi-scenario examination including moving object trajectory prediction, traffic flow prediction, driving scene prediction, and human motion capture.

1. Introduction

Predicting the future is an innate ability possessed by humans, making it a challenging task for machines due to the complex inner laws of the chaotic world. Spatiotemporal predictive learning as a data-driven approach generates future sequences based on historical sequences, with extensive applications including weather forecasting [34, 34], human motion forecasting [2, 42], traffic flow prediction [12, 48], representation learning [24, 29], and vision-based predictive control [13, 19]. In contrast to supervised models that require annotated data, spatiotemporal predictive models can uncover complex spatial and temporal correlations in a self-supervised manner using massive unlabeled data. Spatiotemporal data as the most accessible resource, these methodologies offer potential as unsupervised pre-training paradigms for universal representation learning [4, 32, 38, 39].



(a) Temporal Attention

(b) Spatial Attention

(c) Channel Attention

Figure 1. (a) Temporal attention allows cross-frame interaction tokens to enable long-term modeling. (b) Spatial attention partitions the spatial tokens into global grids and performs the unshuffle operation to implement global spatial interaction. (c) Channel attention is performed in each channel group with linear computational effort. In this work, we alternately use three types of attention to learn short- and long-range spatiotemporal information.

Struggling with the inherent complexity and randomness of future events, spatiotemporal predictive learning has progressively evolved into two approaches, *recurrent-based* and *recurrent-free* frameworks shown in Figure 2. The recurrent-based methods [5, 47, 48] dominate the task due to their superior temporal modeling ability. Many mainstream models [40, 47] with stacked recurrent units capture the temporal dependencies. Inspired by the success

[◇]Corresponding author.

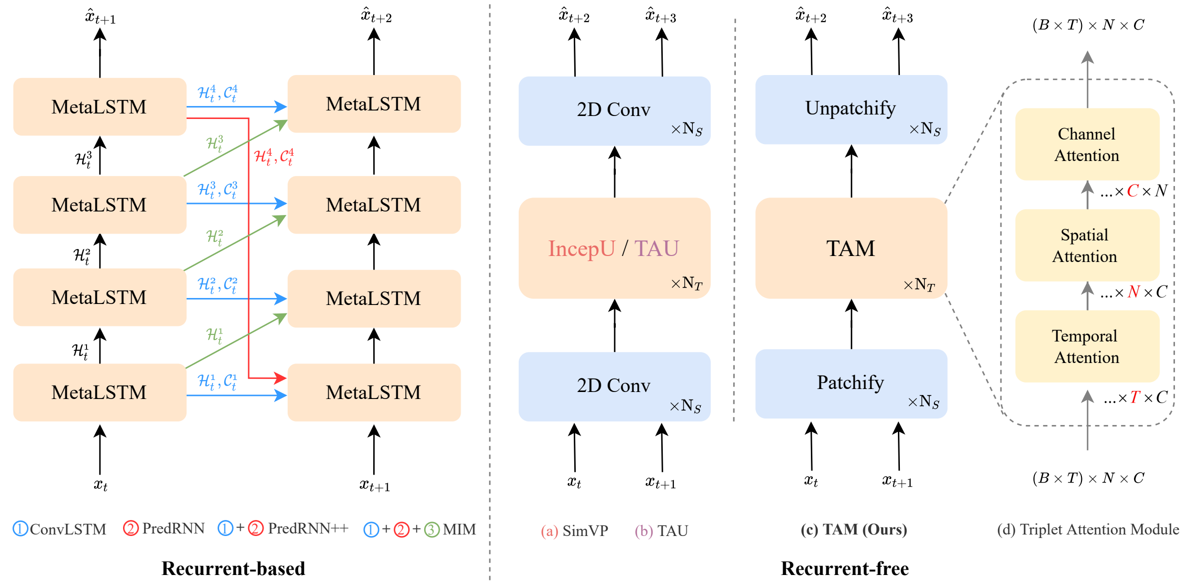


Figure 2. Two typical spatiotemporal predictive learning frameworks. The recurrent-based methods extract spatiotemporal dependencies by cooperating with recurrent units MetaLSTM, and the information transmission between these units. The recurrent-free such as (a) SimVP and (b) TAU extract spatiotemporal features through Inception and temporal attention unit. In contrast, Our approach implements attention mechanisms in each dimension to learn more sophisticated spatiotemporal dependencies.

of long short-term memory (LSTM) networks [21] in sequential modeling, ConvLSTM [35], PredRNN [46], PredRNN++ [44], and MIM [48] propose various LSTM variants, called MetaLSTM, such as ConvLSTM, ST-LSTM, Causal LSTM, and MIM-LSTM. Thus we abstract the general framework of recurrent-based models on the left side of Figure 2, which consists of two main parts: (i) various LSTM variants MetaLSTM; (ii) mode of feature information transmission across different time steps. While the recurrent-based framework is superior in predictive performance, non-parallelizable and computational inefficiency limits its further application. Recently, recurrent-free methods [22, 37, 38] with the parallelizable advantage have been proposed for spatiotemporal learning. As shown in Figure 2(a)(b), we demonstrated the recurrent-free framework representing SimVP [14] and TAU [38], which also consists of two main parts: (i) spatial encoder-decoder; (ii) latent feature spatiotemporal translator. Despite greater computational efficiency, the above methods still have performance gaps in some scenarios compared to the recurrent-based model due to the irrobust modeling of intra- and inter-frame variations.

In this work, we present an innovative triplet attention mechanism that is able to learn short- and long-range sophisticated spatiotemporal dependencies while maintaining computational efficiency. We implemented the Triplet Attention Module (TAM) with an elegant yet simple manner, applying self-attention to the *permutation* of the token matrix, as shown in Figure 1. TAM is decomposed into temporal, spatial, and channel-level attention to capturing temporal and spatial evolution. Specifically, temporal attention modeling inter-frame dynamics, while spatial and channel attention modeling intra-frame static features. We improve spatiotemporal prediction learning by replacing the dominant recurrent unit with a parallelizable pure attention framework. Combining the advantages of a parallelizable framework with the power of Transformer, we implement our proposed TAM blocks and surprisingly find the derived model achieves state-of-the-art in an extensive multi-scenario prediction, including synthetic moving object trajectory prediction, traffic flow prediction, driving scene prediction, and human motion capture. We outline our key contributions as follows:

- We propose the novel *Triplet Attention Transformer* for spatiotemporal predictive learning, which seamlessly integrates intra- and inter-frame feature interaction to obtain powerful representation ability.
- We propose a parallelizable Triplet Attention Module (TAM), which enables models to learn complex short-term and long-term spatiotemporal dependencies through alternating use of *temporal*, *spatial*, and *channel-level* attention.
- We conduct extensive experiments that outperform existing recurrent-based and recurrent-free networks,

achieving state-of-the-art results on Kitti&Caltech, Human3.6M, TaxiBJ, and Moving MNIST datasets.

2. Related Work

Self-Supervised Learning. Despite the notable strides made with supervised learning methods on large labeled datasets, the limited labeled data constrains artificial intelligence development. In contrast, self-supervised learning, using plentiful unlabeled data, offers a promising route toward achieving human-level intelligence. Self-supervised learning creates guiding signals from the data itself via pretext tasks, enabling models to learn data representation. Early visual self-supervised tasks involved colorization [51], inpainting [33], rotation [16], and jigsaw [31]. Contrastive learning [17, 43, 49], while dominant, has limitations on small datasets due to its pair-making process. Masked reconstruction learning [7, 20, 25], which predicts hidden parts from visible ones, is successful in natural language processing but challenging in visual tasks. In contrast to image-level methods, spatiotemporal predictive learning, a burgeoning self-supervised approach, emphasizes video-level information. It predicts upcoming frames by learning from previous ones, thus allowing the model to efficiently segregate foreground and background based on inherent motion dynamics.

Spatiotemporal Predictive Learning. Recent strides in recurrent-based models have provided valuable insights into spatiotemporal predictive learning. ConvLSTM [35], a pioneering work, integrates convolutional networks into LSTM architecture. PredRNN [46] proposes a spatio-temporal LSTM (ST-LSTM) based on vanilla ConvLSTM modules to model spatial and temporal variations. PredRNN++ [44] proposes a Casual-LSTM to connect spatial and temporal memories and a gradient highway unit to mitigate the gradient vanishing. MIM [48] using differential information between hidden states for better non-stationarity handling. E3D-LSTM [45] incorporating 3D convolutions into LSTM architecture. PredRNNv2 [47] proposes a curriculum learning strategy and memory decoupling loss for enhanced performance. MAU [5] designs a motion-aware unit to capture motion information. SwinLSTM [40] integrates the Swin Transformer [27] module into the LSTM architecture for better spatiotemporal modeling. Recently, recurrent-free models have achieved superior performance with the advantage of parallelization. SimVP [14], a seminal work, applies blocks of Inception modules with a UNet architecture to learn the temporal evolution. TAU [38] proposed the temporal attention unit on this basis to capture time evolution. DMVFN [22] proposes a dynamic multi-scale voxel flow network to achieve better prediction performance. Although these recurrent-free methods have achieved great success, their performance is still inferior to recurrent-based in certain scenarios. To this end, we use pure parallelizable attention mechanisms to learn more sophisticated spatiotemporal dependencies.

Vision Transformer. Vision Transformer [10] (ViT) demonstrates exceptional performance across various vision tasks. To enhance its efficiency and effectiveness in image classification tasks, a series of ViT-based approaches have been proposed. Swin Transformer [27] employs local attention windows and implements shift operations to augment window-based interactions. DaViT [8] introduces a dual self-attention mechanism aimed at capturing global context with linear computational complexity. Due to the remarkable performance of ViT, researchers are now using them to understand video content. Uniformer [26] organically unifies convolution and self-attention to solve local redundancy and global dependency. TimeSformer [3] and ViViT [1] explore separate strategies for temporal and spatial attention, achieving excellent outcomes. MViT [11] introduces multiscale features for video sequences, and Video Swin Transformer [28] adapts the model to 3D settings. However, most existing models concentrate on video classification and works about video prediction using ViT are still limited. Is there a solution that combines the strengths of recurrent-based and recurrent-free architecture and takes advantage of the high performance of ViT? Therefore, we propose a triplet attention transformer for efficient spatiotemporal predictive learning.

3. Preliminaries

3.1. Problem Definition

Given $X_{in}^{t:T} = \{X_t, \dots, X_T\}$, the objective is to predict the most reasonable sequences of length T' in the future, denoted as $X_{out}^{T+1:T+T'} = \{\widehat{X}_{T+1}, \dots, \widehat{X}_{T+T'}\}$. We represent the spatiotemporal sequences as a four-dimensional tensor, $i.e.$, $X_{in}^{t:T} \in \mathbb{R}^{T\times C\times H\times W}$, where C, T, H, and W denote channel, temporal or frames, height and width, respectively. The model with learnable parameters θ learns a mapping $\mathcal{F}_\theta : \mathcal{X}_{in}^{t:T} \mapsto \mathcal{X}_{out}^{T+1:T+T'}$ by exploring spatiotemporal dependencies. Concretely, we use the stochastic gradient descent algorithm to learn the mapping $\mathcal{F}_\theta$ and find a set of parameters $\theta^\star$, which minimize the difference between the prediction and the ground-truth, the optimal parameters $\theta^\star$ are:

$$\theta^\star = \arg\min_\theta \mathcal{L}\left(\mathcal{F}_\theta\left(X_{in}^{t:T}\right), X_{out}^{T+1:T+T'}\right), \qquad (1)$$

where $\mathcal{L}$ denote a loss function. In this paper, we adopt the vanilla Mean Squared Error (MSE) as our loss metric.

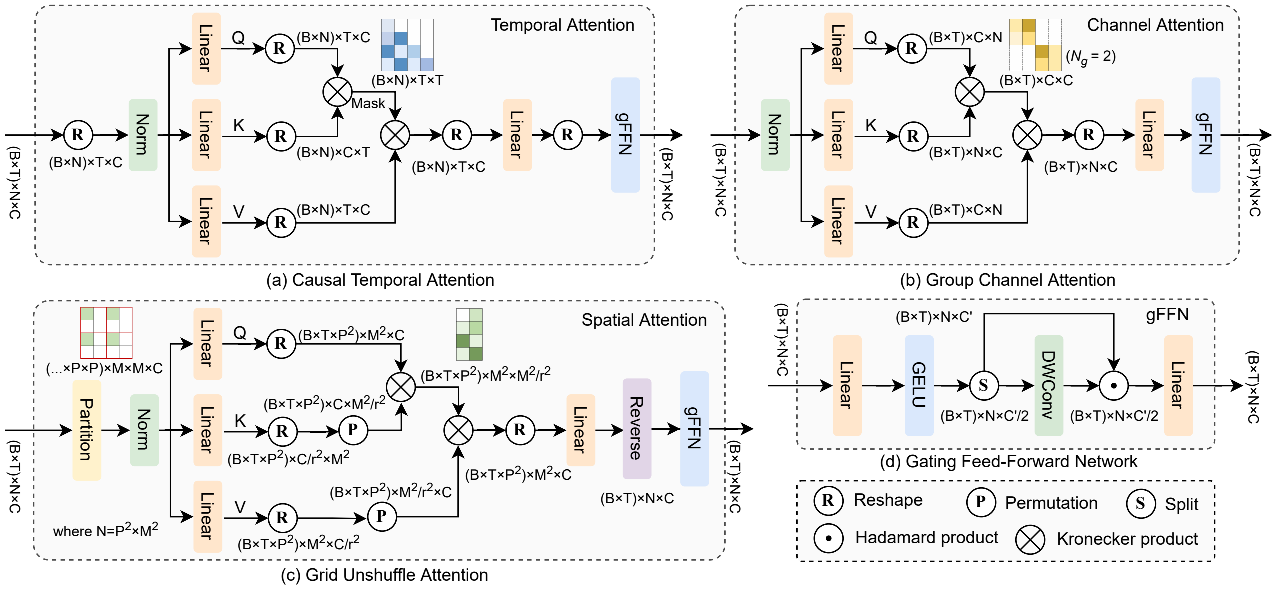


Figure 3. The detailed architecture for triplet attention module. It contains three attention blocks: (a) Causal temporal attention, (b) Group channel attention, and (c) Grid unshuffle attention. By alternately using the three types of attention, our model enjoys the benefit of capturing both spatial dependency and temporal variation. (d) Gating feed-forward network reduces redundant information in channels.

3.2. Self-Attention Mechanism

Assume a visual feature with dimension $\mathbb{R}^{N\times C}$, where N is the number of total patches and C is the number of total channels. Simply applying the standard global self-attention leads to quadratic complexity about input tokens. It is defined as:

$$\begin{aligned}\mathcal{A}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \operatorname{Concat}\left(\text{head}_1,\ldots,\text{head}_{N_h}\right),\\ \text{where head}_i &= \operatorname{Attention}\left(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i\right),\\ &= \operatorname{Softmax}\left[\frac{\mathbf{Q}_i\left(\mathbf{K}_i\right)^{\mathrm{T}}}{\sqrt{C_h}}\right]\mathbf{V}_i,\end{aligned} \tag{2}$$

where $\mathbf{Q}_i = X_i\mathbf{W}_i^Q, \mathbf{K}_i = X_i\mathbf{W}_i^K$, and $\mathbf{V}_i = X_i\mathbf{W}_i^V$ are $\mathbb{R}^{N\times C_h}$ dimensional visual tokens with N_h heads, X_i denotes the i_{th} head of the input tokens and $\mathbf{W}_i$ denotes the projection weights of the i_{th} head for $\mathbf{Q},\mathbf{K},\mathbf{V}$, and $C = C_h * N_h$. Please note that we omit the output projection $\mathbf{W}^O$. It's noteworthy that, due to potential large values of N (*e.g.*, 64×64), the computational implications can be significant.

In this paper, we alternatively arrange causal temporal attention, grid unshuffle attention, and group channel attention to learn more sophisticated spatiotemporal dependencies with less complexity, as shown in Figure 3.

4. Proposed Method

We approach the concept of self-attention from an alternative perspective, proposing a Triplet Attention Module (TAM) that integrates temporal, spatial, and channel-level attention for optimized spatiotemporal predictive learning. Striving for simplicity, the model follows the general framework in Figure 2(c), incorporates the *patchify* and *unpatchify* module which comprises vanilla 2D convolutional and transposed convolutional layers. TAM leverages the self-attention mechanism across varying dimensions, as outlined in Figure 1, we introduce three distinct attention modules: *Causal Temporal Attention*, *Grid Unshuffle Attention*, and *Group Channel Attention*. In the middle layer, the repeated stacking of TAM facilitates the learning of both short-term and long-term complex spatiotemporal dependencies. Therefore, a comprehensive discourse on these modules is provided subsequently.

4.1. Causal Temporal Attention

Previous vision-based self-attention [27, 41, 52], tokens have been defined using pixels or patches, emphasizing spatial dimensions. Instead of spatial attention, we apply attention mechanisms on the temporal-level tokens to capture long-term dependencies. This allows temporal tokens to interact with inter-frame information more efficiently. Although our approach employs a non-autoregressive framework, it can be easily extended to parallelizable autoregression compared to non-parallelizable recurrent-based models. Specifically, we achieve this by masking out (setting to $-\infty$) all values of the upper triangle of the attention matrix to prevent previous frames from seeing subsequent frames, as shown in Figure 3(a).

Simple permutation of feature dimensions can obtain vanilla temporal-level attention. Formally, let T denote the number of frames, N the number of patches, and C the number of channels. Therefore, the token of each frame is designed to interact across other frames. It is defined as:

$$\begin{aligned} \mathcal{A}_{temporal}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \{\mathcal{A}\left(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i\right)\}_{i=0}^{N}, \\ \mathcal{A}\left(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i\right) &= \text{Softmax}\left[\frac{\mathcal{M}(\mathbf{Q}_i^{\mathrm{T}}\mathbf{K}_i)}{\sqrt{C_k}}\right]\mathbf{V}_i^{T}, \end{aligned} \tag{3}$$

where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{T \times C}$ are frame-wise image-level queries, keys, and values, and $\mathcal{M}$ represents the mask operation. The aforementioned equation can be adapted to a multi-head version by dividing the channels into several groups. More details are shown in Figure 3(a).

4.2. Grid Unshuffle Attention

To address the quadratic complexity with the number of input tokens in ViT, we adopted the approach from prior research [30, 41] involving gridded feature maps to aggregate global tokens. A smaller grid size M will result in a gridding effect, Figure 4 illustrates our use of the unshuffle operation to permutate the spatial token to the channel token for an expanded grid size M. Specifically, as depicted in Figure 3(c), for an input feature map $X \in \mathbb{R}^{N \times C}$ and an unshuffle factor r, we partition X into $M \times M$ non-overlapping windows. Subsequently, we gather tokens at identical locations in each window, denoted as $X_p \in \mathbb{R}^{P^2 \times M^2 \times C}$ where $P \times P$ represents the total tokens per window and $N = P^2 \times M^2$. We then employ three linear layers, $\mathbf{W}_i^Q$, $\mathbf{W}_i^K$, and $\mathbf{W}_i^V$, to obtain $\mathbf{Q}_i$, $\mathbf{K}_i$, and $\mathbf{V}_i$:

$$\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i = \mathbf{W}_i^Q(X_p), \mathbf{W}_i^K(X_p), \mathbf{W}_i^V(X_p), \tag{4}$$

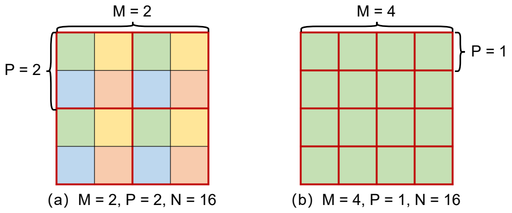
Here, $\mathbf{Q}_i$ maintains its channel dimension, whereas $\mathbf{W}_i^K$ and $\mathbf{W}_i^V$ reduce it to C/r^2, resulting in $\mathbf{K}_i \in \mathbb{R}^{P^2 M^2 \times C/r^2}$ and $\mathbf{V}_i \in \mathbb{R}^{P^2 M^2 \times C/r^2}$. Spatial tokens in these matrices are permuted to channel tokens, producing $\mathbf{K}_i^p \in \mathbb{R}^{C \times M^2/r^2}$ and $\mathbf{V}_i^p \in \mathbb{R}^{M^2/r^2 \times C}$. Using $\mathbf{Q}_i$ with these permuted tokens, we execute the self-attention operation, enabling larger grid sizes (e.g., 24×24) with fewer computations than 8×8, yet achieving superior performance. It is defined as:

$$\begin{aligned} \mathcal{A}_{spatial}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \{\mathcal{A}\left(\mathbf{Q}_i,\mathbf{K}_i^p,\mathbf{V}_i^p\right)\}_{i=0}^{P}, \\ \mathcal{A}\left(\mathbf{Q}_i,\mathbf{K}_i^p,\mathbf{V}_i^p\right) &= \text{Softmax}\left[\frac{(\mathbf{Q}_i(\mathbf{K}_i^p)^{\mathrm{T}}}{\sqrt{C_k}} + \mathbf{B}\right]\mathbf{V}_i^p, \end{aligned} \tag{5}$$

where $\mathbf{B}$ is the relative position embedding.

4.3. Group Channel Attention

Channel Attention is similar to the above modules, as illustrated in Figure 1(c), we employ self-attention in the channel dimension. Notably, in many scenarios, the number of channels will be higher ($e.g.$, $C = 256$). To mitigate the inherent quadratic complexity of self-attention concerning the channel dimension, we group channels into multiple groups and perform self-attention within each group. Formally, let N_g denote the number of groups and C_g the number of channels in each group, we have $C = N_g * C_g$. More details are shown in Figure 3(b). It is defined as:



(a) M = 2, P = 2, N = 16

(b) M = 4, P = 1, N = 16

Figure 4. Compressing channels permutate spatial tokens into channel tokens allowing larger grid size M with less computation.

$$\begin{aligned} \mathcal{A}_{channel}(\mathbf{Q},\mathbf{K},\mathbf{V}) &= \left\{\mathcal{A}\left(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i\right)^T\right\}_{i=0}^{N_g}, \\ \mathcal{A}\left(\mathbf{Q}_i,\mathbf{K}_i,\mathbf{V}_i\right) &= \text{Softmax}\left[\frac{\mathbf{Q}_i^{\mathrm{T}}\mathbf{K}_i}{\sqrt{C_g}}\right]\mathbf{V}_i^{T}. \end{aligned} \tag{6}$$

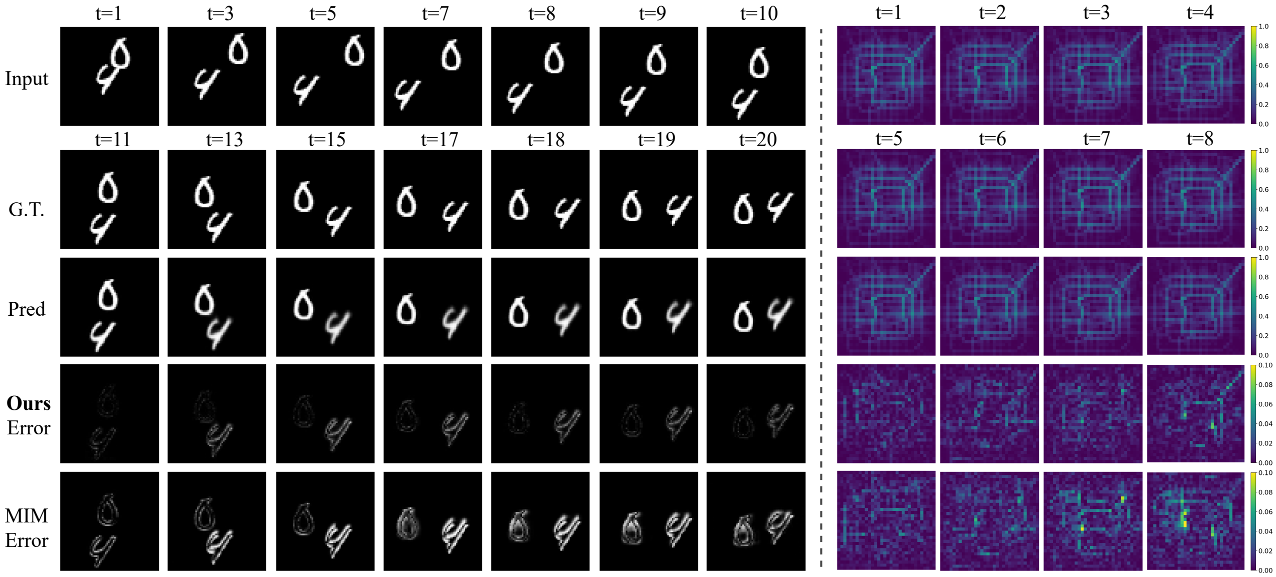
where $\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i \in \mathbb{R}^{N \times C_g}$ as channel-grouped queries, keys, and values. To accommodate frames with different sizes, the projection layers $\mathbf{W}$ remain performed along the channel dimension. We also use conditional positional encoding [6] (CPE) to provide location information.

5. Experiments

Multi-Scenario Examination. Our model is quantitatively evaluated across expansive real-world scenarios with diverse scales, including traffic flow prediction, driving scene prediction, and human motion capture. For synthetic data scenarios such as Moving MNIST [36], we also offer comprehensive experiments. Each dataset gathers from various domains, from micro to macro scales. The details of dataset statistics are shown in Table 1. For more experiments, please refer to the Supplementary Materials.

- **Synthetic Moving Object Trajectory Prediction.** The *Moving MNIST* dataset [36] as a foundational benchmark that has been widely employed in various studies. This dataset is comprised of video sequences in which two digits traverse across a frame of dimensions 64×64 pixels. For each digit, its velocity is determined by two factors: (i) a direction randomly chosen from a unit circle; (ii) a magnitude arbitrarily selected from a predefined range.

- **Traffic Flow Prediction.** Efficient traffic governance and public safety depend on accurate crowd dynamics



(a) Moving MNIST

(b) TaxiBJ

Figure 5. Qualitative visualizations on (a) Moving MNIST and (b) TaxiBJ, where prediction error = |ground truth - prediction|.

prediction. We employ the *TaxiBJ* dataset [50], involving taxi GPS trajectories from Beijing bifurcated into two channels: inflow and outflow. This dataset has a 30-minute interval and a spatial granularity of 32 × 32. Our approaches to data preprocessing, model training, and performance assessment are congruent with protocols by PredNet [18] and MIM [48].

- **Driving Scene Prediction.** In autonomous driving, predicting future dynamics is critically important in complex and non-stationary environments. We employ two datasets for evaluation purposes: *KITTI* [15] extensively utilized in the fields of autonomous driving and robotics; *Caltech Pedestrian* [9] specializes in pedestrian detection. We train our model on the *KITTI* dataset and evaluate performance on the *Caltech Pedestrian* benchmark.
- **Human Motion Capture.** Predicting human motion remains a formidable challenge due to the considerable variability across individual behaviors and actions. In our study, we utilize the *Human3.6M* [23] dataset, encompassing high-resolution motion capture videos. Following previous work settings [18], we employ four observed frames to predict the subsequent four frames.

Evaluation Metrics. We evaluate the performance of the proposed model using various metrics. For pixel-wise error, we consider mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE). Structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR) are used for similarity evaluation.

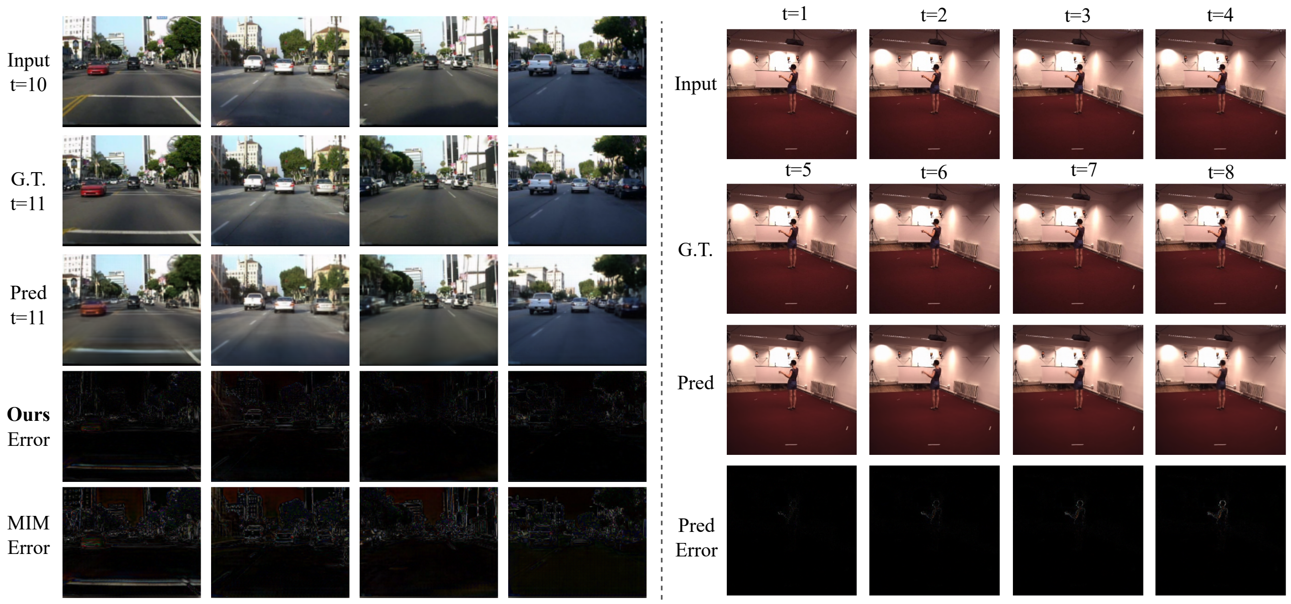
Dataset	Train	Test	C	H	W	T	T'
Kitti&Caltech	3,160	3,095	3	128	160	10	1
Human3.6M	73,404	8,582	3	256	256	4	4
TaxiBJ	20,461	500	2	32	32	4	4
Moving MNIST	10,000	10,000	1	64	64	10	10

Table 1. The details of dataset statistics. We detail the number of samples, the input frames denoted as T, and the predicted frames represented as T' for both the training and testing subsets.

Implementation Details. We use the PyTorch framework on a single NVIDIA-V100 GPU for our proposed method. The model trains with mini-batches of 16 video sequences using the AdamW optimizer, the OneCycle learning rate scheduler, and a weight decay of $5e^{-2}$. The learning rate, selected from $\{1e^{-2}, 5e^{-3}, 1e^{-3}\}$, ensures stable training. We employ stochastic depth as regularization.

5.1. Synthetic Moving Object Trajectory Prediction

Moving MNIST. This dataset is a standard benchmark for evaluating spatiotemporal predictive learning methods. We compare our proposed approach with various recent strong baselines. The quantitative results are detailed in Table 2, and visualizations of the predictions can be found in Figure 5(a). Notably, our method exceeds all baselines under four separate metrics. Compared to ConvLSTM [35], our method reduces the MSE from 103.3 to 17.55 and increases the SSIM from 0.707 to 0.966. In contrast to MIM [48], our method can accurately predict motion trajectories and appearances of two digits. We also tried to experiment with autoregressive (w/ AR) methods of the recurrent-based ar-



(a) Kitti&Caltech

(b) Human3.6M

Figure 6. Qualitative visualizations on (a) Kitti&Caltech and (b) Human3.6M, where prediction error = |ground truth - prediction|.

chitectures, finding that while it improved prediction quality, more time was needed to complete the training.

		Moving MNIST (10 → 10 frames)			
Method	Reference	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑
ConvLSTM [35]	NIPS'2015	103.3	182.9	0.707	16.17
PredRNN [46]	NIPS'2017	56.8	126.1	0.867	19.12
PredRNN++ [44]	ICML'2018	46.5	106.8	0.898	20.11
MIM [48]	CVPR'2019	44.2	101.1	0.910	20.31
E3D-LSTM [45]	ICLR'2019	41.3	87.2	0.910	20.70
MAU [5]	NIPS'2021	27.6	86.5	0.937	22.59
PredRNNv2 [47]	TPAMI'2022	48.4	129.8	0.891	20.12
SimVP [14]	CVPR'2022	23.8	68.9	0.948	23.19
TAU [38]	CVPR'2023	19.8	60.3	0.957	24.53
DMVFN [22]	CVPR'2023	123.6	179.9	0.814	16.15
Ours	-	17.55	59.81	0.960	25.08
Ours w/ AR	-	**15.68**	**51.85**	**0.966**	**25.71**

Table 2. Quantitative results on the Moving MNIST dataset.

5.2. Traffic Flow Prediction

Taxibj. Traffic flow prediction presents significant challenges due to the unpredictability introduced by human behavior. Our method is evaluated using the TaxiBJ dataset [50], which embodies the complex nature inherent in real-world traffic systems. The complexity of road networks and nonlinear temporal behaviors limit the efficacy of traditional forecasting methods.

Table 3 reports the quantitative results, while Figure 5(b) offers qualitative visualizations. To optimize the visual interpretation, the error scale is limited to 0.1 and focuses solely on the inflow case. Despite minor deviations between observed and future data frames, our model consistently yields precise forecasts compared to recurrent-based methods. Owing to the robust spatiotemporal relationships captured by the triplet attention module, our methodology sets new benchmarks across all evaluation metrics, suggesting its suitability for application in traffic flow prediction.

		TaxiBJ (4 → 4 frames)			
Method	Reference	MSE × 100↓	MAE ↓	SSIM ↑	PSNR ↑
ConvLSTM [35]	NIPS'2015	48.5	17.7	0.978	37.38
PredRNN [46]	NIPS'2017	46.4	17.1	0.971	38.52
PredRNN++ [44]	ICML'2018	44.8	16.9	0.977	38.71
MIM [48]	CVPR'2019	42.9	16.6	0.971	38.71
E3D-LSTM [45]	ICLR'2019	43.2	16.9	0.979	38.75
PhyDNet [18]	CVPR'2020	41.9	16.2	0.982	39.18
SimVP [14]	CVPR'2022	41.4	16.2	0.982	39.17
PredRNNv2 [47]	TPAMI'2022	38.3	15.6	0.983	39.38
TAU [38]	CVPR'2023	34.4	15.6	0.983	39.50
SwinLSTM [40]	ICCV'2023	43.1	17.3	0.977	38.71
Ours	-	**31.3**	**15.1**	**0.984**	**39.67**

Table 3. Quantitative results in the TaxiBJ dataset.

5.3. Driving Scene Prediction

Kitti&Caltech. The ability to generalize is important in artificial intelligence. Traditional supervised learning often has limitations when applied to diverse domains. In contrast, self-supervised learning methods, such as contrastive learning and masked reconstruction learning, aim to learn robust representations from unlabeled data. These models then evaluate generalization ability through downstream tasks. In this paper, we evaluated this ability across different datasets, where we train our model on the KITTI [15] and then evaluate its performance on the Caltech Pedestrian [9].

Figure 6 presents our qualitative visualizations, while Table 4 offers the quantitative results. Remarkably, our method not only surpasses all recurrent-based approaches but also establishes new state-of-the-art results. It can be seen from the prediction errors in the last two rows of Figure 6(a), that our model effectively predicts both lane lines and distant vehicles. Given its consistent accuracy in dealing with variations in lighting and lane lines, our approach shows promise for application in autonomous vehicles.

Method	Kitti&Caltech (10 → 1 frames) Reference	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑
ConvLSTM [35]	NIPS'2015	139.6	1583.3	0.9345	27.46
PredRNN [46]	NIPS'2017	130.4	1525.5	0.9374	27.81
PredRNN++ [44]	ICML'2018	129.6	1507.7	0.9453	27.89
MIM [48]	CVPR'2019	127.4	1476.5	0.9461	27.98
E3D-LSTM [45]	ICLR'2019	200.6	1946.2	0.9047	25.45
PhyDNet [18]	CVPR'2020	312.2	2754.8	0.8615	23.26
MAU [5]	NIPS'2021	177.8	1800.4	0.9176	26.14
SimVP [14]	CVPR'2022	160.2	1690.8	0.9338	26.81
PredRNNv2 [47]	TPAMI'2022	147.8	1610.5	0.9330	27.12
TAU [38]	CVPR'2023	131.1	1507.8	0.9456	27.83
DMVFN [22]	CVPR'2023	183.9	1531.1	0.9314	26.78
Ours	-	**122.9**	**1416.2**	**0.9469**	**28.18**

Table 4. Quantitative results in Kitti&Caltech dataset.

5.4. Human Motion Capture

Human3.6M. Predicting human motion is challenging due to both the need for high-resolution forecasting and the complexity introduced by human unpredictability. To provide a comprehensive evaluation from multiple perspectives, we employ MSE, MAE, SSIM, and PSNR as metrics. Table 5 provides qualitative results, and it can be seen that our method consistently outperforms the recurrent-based methods and establishes a strong baseline. We also present the visualization in Figure 6(b), where the smaller prediction error reveals that our method can handle real-world dynamic scenarios.

5.5. Ablation Study

Triplet attention layout. We tested four configurations for our triplet attention mechanism: (i) temporal attention first; (ii) spatial attention first; (iii) channel attention first; and (iv) triplet attention parallel. Table 6 results show similar performance across all settings, with a slight edge for 'temporal attention first'.

Effects of different attention. We evaluate the contribution of different attentions by removing certain attention to compare performance on the Human3.6M [23] dataset. Table 6 shows the results, it can be seen that temporal attention is relatively important as it models inter-frame dynamics, and the other two attentions model intra-frame static.

Method	Human3.6M (4 → 4 frames) Reference	MSE ↓	MAE ↓	SSIM ↑	PSNR ↑
ConvLSTM [35]	NIPS'2015	125.5	1566.7	0.9813	33.40
PredRNN [46]	NIPS'2017	113.2	1458.3	0.9831	33.94
PredRNN++ [44]	ICML'2018	111.3	1454.4	0.9832	33.92
MIM [48]	CVPR'2019	112.1	1467.1	0.9829	33.97
E3D-LSTM [45]	ICLR'2019	143.3	1442.5	0.9803	32.52
PhyDNet [18]	CVPR'2020	125.7	1614.7	0.9804	33.05
MAU [5]	NIPS'2021	127.3	1577.0	0.9812	33.33
SimVP [14]	CVPR'2022	115.8	1511.5	0.9822	33.73
PredRNNv2 [47]	TPAMI'2022	114.9	1484.7	0.9827	33.84
TAU [38]	CVPR'2023	113.3	1390.7	**0.9839**	34.03
Ours	-	**108.4**	**1389.1**	**0.9839**	**34.18**

Table 5. Quantitative results in Human3.6M dataset.

Method	SSIM↑	PSNR↑
Temporal Attention First	**0.9839**	**34.18**
Spatial Attention First	0.9826	34.10
Channel Attention First	0.9824	34.07
Triplet Attention Parallel	0.9804	33.12
Triplet Attention Module	0.9839	34.18
- Temporal Attention	0.9794 **(-0.0045)**	32.77 **(-1.41)**
- Spatial Attention	0.9809 (-0.0030)	33.26 (-0.92)
- Channel Attention	0.9813 (-0.0026)	33.55 (-0.63)

Table 6. Ablation study in Human3.6M dataset.

6. Conclusion

This work introduces a novel triplet attention mechanism comprising causal temporal attention, grid unshuffle attention, and group channel attention. This mechanism effectively learns short and long-range spatiotemporal dependencies while maintaining computational parallelism. The three self-attentions are complementary: (i) temporal attention captures temporal dependence due to the abstract representations in each temporal token; (ii) spatial and channel attention together refine intra-frame representation via fine-grained interactions across spatial and channel dimensions. Extensive validation across multiple scenarios demonstrates the superior performance of our method. In sum, our approach highlights the importance of both intra-frame and inter-frame variations and provides a novel perspective on spatiotemporal predictive learning.

Acknowledgements

This work is supported by the CCF-Lenovo Bule Ocean Research Fund.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[2] Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H Campbell, and Sergey Levine. Stochastic variational video prediction. *arXiv preprint arXiv:1710.11252*, 2017. 1

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3

[4] 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

[5] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, Yan Ye, Xiang Xinguang, and Wen Gao. Mau: A motion-aware unit for video prediction and beyond. *Advances in Neural Information Processing Systems*, 34:26950–26962, 2021. 1, 3, 7, 8

[6] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021. 5

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ACL*, 2019. 3

[8] Mingyu Ding, Bin Xiao, Noel Codella, Ping Luo, Jingdong Wang, and Lu Yuan. Davit: Dual attention vision transformers. In *European Conference on Computer Vision*, pages 74–92. Springer, 2022. 3

[9] 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. 6, 8

[10] 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. *arXiv preprint arXiv:2010.11929*, 2020. 3

[11] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6824–6835, 2021. 3

[12] 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

[13] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016. 1

[14] 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*, pages 3170–3180, 2022. 2, 3, 7, 8

[15] 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. 6, 8

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *ICLR*, 2018. 3

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 3

[18] Vincent Le Guen et al. 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. 6, 7, 8

[19] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. *CVPR*, 2022. 1

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 3

[21] Sepp Hochreiter et al. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[22] Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, and Shuchang Zhou. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2023. 2, 3, 7, 8

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 6, 8

[24] 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

[25] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ACL*, 2020. 3

[26] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[27] 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 *ICCV*, pages 10012–10022, 2021. 3, 4

[28] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 3

[29] William Lotter et al. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016. 1

[30] Sachin Mehta and Mohammad Rastegari. Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv preprint arXiv:2110.02178*, 2021. 5

[31] 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. 3

[32] 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*, 44(6):2806–2826, 2020. 1

[33] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *CVPR*, pages 2536–2544, 2016. 3

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

[35] 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, 6, 7, 8

[36] 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. 5

[37] Cheng Tan, Zhangyang Gao, Siyuan Li, and Stan Z Li. Simvp: Towards simple yet powerful spatiotemporal predictive learning. *arXiv preprint arXiv:2211.12509*, 2022. 2

[38] Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, and Stan Z Li. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18770–18782, 2023. 1, 2, 3, 7, 8

[39] Cheng Tan, Siyuan Li, Zhangyang Gao, Wenfei Guan, Zedong Wang, Zicheng Liu, Lirong Wu, and Stan Z Li. Openstl: A comprehensive benchmark of spatio-temporal predictive learning. *arXiv preprint arXiv:2306.11249*, 2023. 1

[40] Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. *arXiv preprint arXiv:2308.09891*, 2023. 1, 3, 7

[41] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 4, 5

[42] 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

[43] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, pages 9929–9939. PMLR, 2020. 3

[44] 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, 7, 8

[45] 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. 3, 7, 8

[46] 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, 7, 8

[47] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, S Yu Philip, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022. 1, 3, 7, 8

[48] 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, 2, 3, 6, 7, 8

[49] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *ICML*, pages 12310–12320. PMLR, 2021. 3

[50] Junbo Zhang et al. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 6, 7

[51] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 3

[52] Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, and Rynson WH Lau. Biformer: Vision transformer with bi-level routing attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10323–10333, 2023. 4