Electronic
Research Archive

https://www.aimspress.com/journal/era

ERA, 32(7): 4321–4339.
DOI: 10.3934/era.2024194
Received: 11 March 2024
Revised: 13 June 2024
Accepted: 20 June 2024
Published: 08 July 2024

Research article

Fully convolutional video prediction network for complex scenarios

Rui Han[1,*], **Shuaiwei Liang**[2], **Fan Yang**[2], **Yong Yang**[1] **and Chen Li**[1]

[1] Electric Power Science Research Institute, State Grid Zhejiang Electric Power Co., Ltd., Hangzhou 310014, China

[2] State Grid Zhejiang Electric Power Co., Ltd., Hangzhou 310014, China

* **Correspondence:** Email: 404468876@qq.com.

Abstract: Traditional predictive models, often used in simpler settings, face issues like high latency and computational demands, especially in complex real-world environments. Recent progress in deep learning has advanced spatiotemporal prediction research, yet challenges persist in general scenarios: (i) Latency and computational load of models; (ii) dynamic nature of real-world environments; (iii) complex motion and monitoring scenes. To overcome these challenges, we introduced a novel spatiotemporal prediction framework. It replaced high-latency recurrent models with fully convolutional ones, improving inference speed. Furthermore, it addressed the dynamic nature of environments with multilevel frequency domain encoders and decoders, facilitating spatial and temporal learning. For complex monitoring scenarios, a large receptive field token mixer spatial-frequency attention units (SAU) and time attention units (TAU) ensured temporal and spatial continuity. This framework outperformed current methods in accuracy and speed on public datasets, showing promising practical applications beyond electricity monitoring.

Keywords: attention mechanism; behavior recognition; fully convolutional network

1. Introduction

With the rapid growth of the video surveillance industry, the network of surveillance systems has expanded significantly year after year [1]. However, there is an acute shortage of surveillance and maintenance personnel [2]. Surveillance and maintenance tasks encompass a wide range of activities characterized by complexity, heavy workloads, tight schedules, harsh working conditions, high-altitude activities, and heavy lifting, presenting a variety of high-risk factors [3, 4]. Unfortunately, the current state of incident detection is not encouraging [5]. The urgent need for early warning to support surveillance operations and maintenance highlights the importance of early prediction of abnormal behavioral patterns through video analysis [6, 7]. This has far-reaching implications for both personnel safety and the stability of the surveillance operations. Human cognitive and perceptual abilities rely on predictive mechanisms that anticipate future events and sensory signals [8–10]. The ability of these mechanisms to rapidly and accurately model and evaluate future events, even when computational power is limited, offers the potential to effectively model the expectation of future events in complex and dynamic environments, promising to reduce the frequency of incidents related to surveillance.

In recent years, the rapid advances in artificial intelligence, especially deep learning, have made significant contributions to the field of spatiotemporal predictive pearning (STL). STL pertains to the challenging task of video prediction [11–13], wherein the goal is to model the distribution of historical spatiotemporal data. This necessitates the acquisition of knowledge about the underlying principles governing our chaotic world [14–16].

After a comprehensive review of prior research on spatiotemporal prediction, it is apparent that the predominant models employed for video prediction tasks are founded on classical recurrent neural networks (RNN). While RNNs excel at modeling time series data, their spatial modeling capabilities are notably lacking [17]. Recognizing this deficiency, recent efforts have sought to bridge the spatial modeling gap in RNNs by devising hybrid architectures that integrate convolutional neural networks (CNNs) [17]. In a pioneering development inspired by the long short-term memory (LSTM) networks within RNNs, [17] introduced the ConvLSTM framework, extending the fully connected (FC) LSTM structure into a convolutional format to enhance precipitation forecasting accuracy. [18] introduced another notable approach, PredRNN, which utilizes spatiotemporal LSTM (ST-LSTM) to capture spatial features and temporal dynamics, achieving highly accurate long-term forecasts. These models achieved significant improvements by reimagining the recurrent units, thus laying the foundation for subsequent work in spatiotemporal learning [19]. [20] amalgamated 3D convolution and LSTM to extract short-term-dependent spatial and motion features, bolstering long-term memory at the temporal level. Moreover, [21] proposed a physically constrained dual-branch structure comprising PhyCell and ConvLSTM to encode prior physical knowledge via the simulation of partial differential equations in the potential space. [22] introduced a reversible neural network architecture, constructing a two-way reversible self-encoder for the acquisition of spatiotemporal insights.

Video prediction tasks find diverse applications in robotic vision planning, traffic flow forecasting, autonomous driving, weather prediction, and surveillance systems [11, 23, 24]. For instance, in robot vision planning, predictive models endowing robots with similar predictive capabilities facilitate multitask planning in complex, dynamic environments. While many industries commonly employ algorithms, including target detection and semantic segmentation [25, 26], to detect violations and issue alerts in real-time video surveillance, hazards in these environments typically manifest after these violations have occurred. Therefore, employing video prediction tasks in surveillance scenarios can significantly enhance the early warning capabilities of surveillance systems, thereby boosting safety and efficiency. Notably, video prediction tasks have yet to be widely adopted in this context, and this paper outlines three key challenges encountered in this context. (i) Model inference latency and computational demands: Addressing model inference latency is pivotal in surveillance scenarios, where any delay could compromise safety. Video prediction models must rapidly and accurately forecast future situations. (ii) Environmental variability in surveillance environments: The real-world surveillance environment is highly complex, influenced by factors such as weather, lighting, environmental noise, and seasonal variations. These changes impact video color and brightness, consequently affecting the model's performance and prediction accuracy. (iii) Complex motions and scenes in video surveillance:

Such surveillance typically involves multiple objects like vehicles, pedestrians, and infrastructure, exhibiting varying states, shapes, and complex motion at different times and angles. Ensuring temporal and spatial continuity is imperative, intensifying the complexity of the video prediction task.

To mitigate model inference latency, prior research has primarily concentrated on directly modeling spatiotemporal relationships using hybrid RNN designs or other networks. However, real-time prediction tasks often entail trade-offs between speed and accuracy when designing recurrent structures. As depicted in Figure 1, this paper reviews previous contributions and abstracts the overarching framework for enhancing spatiotemporal prediction modeling and addressing inference speed. The traditional RNN structure is replaced with the fully convolutional networks (FCN) architecture for modeling spatiotemporal relations, comprising encoder, decoder, and spatiotemporal learning modules. In the context of power scenarios, where sampled video frames exhibit substantial inter-frame differences due to the complex real-world environment influenced by weather, lighting, environmental noise, and other factors, directly modeling the space-time relationship is challenging. To tackle this challenge, frequency domain techniques are harnessed to consider the information regarding overall motion patterns. The DWT decomposes images into approximate and detailed components, enabling lossless reconstruction. A deep CNN is employed to model the time-frequency relationship, especially suited for complex dynamic environments. Therefore, this paper introduces a multilevel 2D DWT encoder-decoder, augmenting the traditional encoder-decoder approach. Moreover, to address complex motion scenarios, a dynamic spatiotemporal attention unit is proposed to dynamically model the temporal and spatial relationships, ensuring temporal-spatial consistency and continuity. Key innovations of this paper include:

- The introduction of a novel spatiotemporal prediction framework that employs a FCN instead of high-latency RNNs. This model computes future video frames in parallel.
- The proposal of multilevel frequency domain codecs, where a wavelet transform encoder establishes time-frequency relationships within and between consecutive video frames, and an inverse wavelet transform decoder (IWT) reconstructs these relationships to future frames in a lossless manner.
- The introduction of the dynamic temporal attention unit (TAU) and spatial attention unit (SAU) to dynamically model the time-frequency relationship in the potential space, ensuring spatio-temporal consistency and continuity.

While recurrent-based methods have excelled in STL, they face challenges in computational efficiency. In addressing the need for parallelizing temporal evolution modeling, we introduce TAU, utilizing visual attention mechanisms without recurrent architecture. Notable prior works like PredCNN [27] utilize pure CNNs for temporal modules, while SimVP [35] employs Inception modules within a UNet architecture for temporal evolution learning. However, we argue that convolutional approaches alone may not adequately capture long-term dependencies. SimVP, while offering a simple baseline, leaves substantial room for further enhancements. Our work replaces the Inception-UNet model with efficient attention modules to improve prediction performance. By employing a straightforward yet powerful attention mechanism [28], our approach not only facilitates parallelization but also captures long-term temporal evolution.

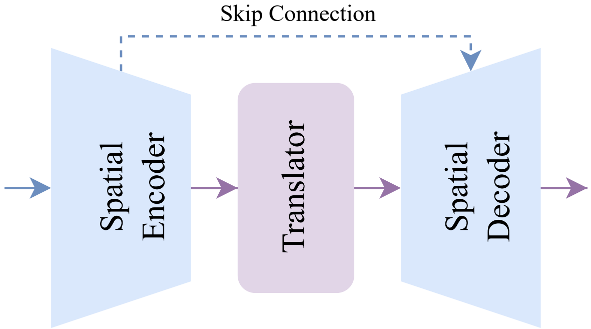

Figure 1. The overview architecture of STU, including spatial encoder, latent translator, and spatial decoder.

2. Problem definition

Given $X_{in}^{t:T}$, representing the input spatiotemporal sequences from time t to T, our objective is to predict future sequences of length T' denoted as $X_{out}^{T+1:T+T'}$. These sequences are treated as four-dimensional tensors, $X_{in}^{t:T} \in \mathbb{R}^{T\times C\times H\times W}$, with C, T, H, and W representing channel, temporal or frame dimension, height, and width, respectively. The model with learnable parameters θ is trained to learn a mapping $\mathcal{F}_\theta : \mathcal{X}_{in}^{t:T} \mapsto \mathcal{X}_{out}^{T+1:T+T'}$ by capturing spatiotemporal dependencies. Specifically, stochastic gradient descent is employed to train the model and determine optimal parameters $\theta^\star$ that minimize the discrepancy between predictions and ground-truth sequences. Mathematically, this is expressed as:

$$\theta^\star = \arg\min_\theta \mathcal{L}\left(\mathcal{F}_\theta\left(X_{in}^{t:T}\right), X_{out}^{T+1:T+T'}\right), \tag{2.1}$$

where $\mathcal{L}$ denotes a loss function, with the mean squared error (MSE) serving as the loss metric in this paper.

3. Proposed methods

Illustrated in Figure 1, our proposed spatio-temporal understanding (STU) framework comprises three main components: an input encoder, a spatiotemporal learning module, and an output decoder. The input encoder transforms individual frames within the input video sequence into a high-dimensional latent space. The spatiotemporal learning module captures both spatial correlations and temporal changes within this latent space. Lastly, the output decoder reconstructs future video frames from the same latent space, offering a comprehensive approach for dynamic video analysis.

Given a batch of input frames $\mathcal{B}$ in $\mathbb{R}^{B\times T\times C\times H\times W}$, with B denoting the batch size, the input and output encoders reshape the tensor into dimensions $(B \times T) \times C \times H \times W$ for subsequent processing. The codec handles each video frame independently and disregards temporal changes. The spatiotemporal learning module reshapes the tensor into dimensions $B \times (T \times C) \times H \times W$, enabling not only intra-frame feature extraction but also modeling of inter-frame temporal variations. This transformation efficiently characterizes spatiotemporal relationships as a set of output potential features, crucial for understanding complex video dynamics.

3.1. Multilevel wavelet transform encoder-decoder

To adapt to the changing dynamics of the power system, this study explores the extraction of spatial multi-scale features from a frequency domain perspective. Since our task is to process video and, thus, we need to analyze images and extract features from them, an effective way of processing images in frequency domain is needed, and that's why we use the 2D wavelet transformation, which performs greatly in this field. The 2D DWT plays a pivotal role in this process. As illustrated in Figure 2, the original image is decomposed into four sub-images, each associated with distinct frequency ranges (f_{LL}, f_{LH}, f_{HL}, and f_{HH}). Here, we give a simple explanation of these four abbreviations, LL, LH, HL and HH.

- LL(low-low) sub-band: This sub-band contains the low-frequency components in both the horizontal and vertical directions. It generally represents the coarse approximation of the signal or image, capturing the overall structure and major features with smooth variations.
- LH(low-high) sub-band: This sub-band contains the low-frequency components in the horizontal direction and the high-frequency components in the vertical direction. It highlights the vertical details and edges in the signal or image.
- HL(high-low) sub-band: This sub-band contains the high-frequency components in the horizontal direction and the low-frequency components in the vertical direction. It highlights the horizontal details and edges in the signal or image.
- HH(high-high) sub-band: This sub-band contains the high-frequency components in both the horizontal and vertical directions. It captures the fine details and textures in the signal or image, including noise and sharp transitions.

This decomposition allows for a detailed analysis of the image's frequency components, essential for tasks like feature extraction and anomaly detection in dynamic systems.

The approximation component of the original image, for example, can be expressed as $x_1 = (f_{LL} \otimes x) \downarrow_2$. This representation offers a compact yet informative summary of the original image's low-frequency components, which are often crucial for understanding the overall structure and trends in the data. The high-frequency components, represented by f_{LH}, f_{HL}, and f_{HH}, capture the finer details and textures of the image, providing a comprehensive view of the spatial characteristics. The nondestructive reconstruction of the original image using the inverse transform, such as $x = \text{IWT}(x_1, x_2, x_3, x_4)$, ensures that the process is reversible, allowing for the original image to be accurately reconstructed from its wavelet components. This feature is particularly important in applications where the integrity and fidelity of the reconstructed image are critical, such as in medical imaging or high-definition video processing.

Furthermore, the sub-images within the four frequency bands produced by the multilevel 2D DWT undergo further decomposition via the DWT, potentially yielding three or more levels of decomposition. This multilevel approach allows for a more granular analysis of the image's frequency components, enabling the extraction of features at various scales and resolutions. Typically, wavelet decomposition requires nonlinear combinations of features, including techniques like soft thresholding and quantization, commonly used in image denoising and compression [29, 30]. These techniques help in reducing noise and compressing the image data without significant loss of quality, making them essential in applications where bandwidth or storage space is limited. To adapt the multi-scale wavelet transformation for video prediction tasks, this paper extends it into an encoder-decoder framework.

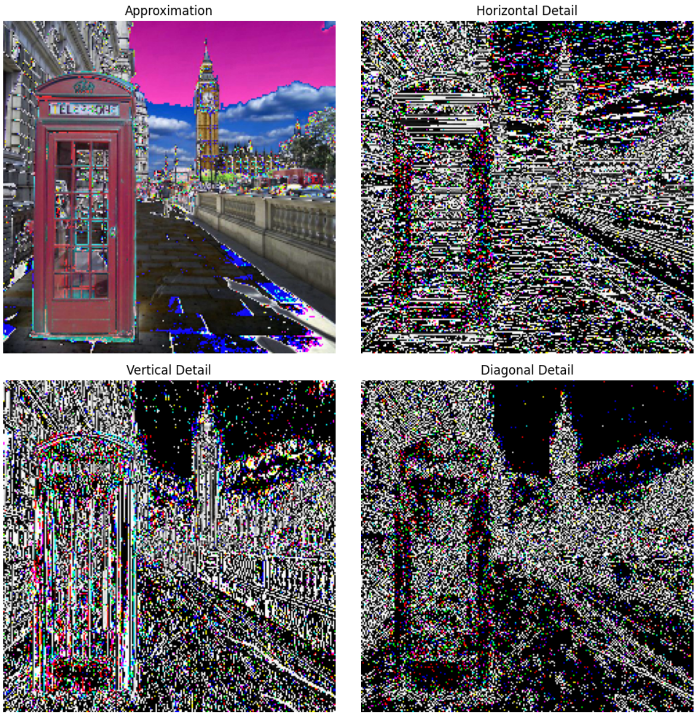

Figure 2. Two-dimensional discrete wavelet transform.

This extension incorporates a convolutional layer, as depicted in Figure 3, facilitating the representation of the coding and decoding process for a video frame. The convolution layer we used in spatial encoder-decoder is 1 × 1 kernel, which is aimed to combine the feature information of different tunnels. That will enhance the network's perception of global features. This process can be mathematically expressed as:

$$z_i = \sigma\left(\mathrm{Nom}\left(\mathrm{Conv}\left(\mathrm{DWT}\left(x_{i-1}\right)\right)\right)\right), \tag{3.1}$$

$$1 \leq i \leq N_e \tag{3.2}$$

$$z_k = \sigma\left(\mathrm{Norm}\left(\mathrm{Conv}\left(\mathrm{IWT}\left(x_{i-1}\right)\right)\right)\right), \tag{3.3}$$

$$N_e + N_t < k \leq 2N_e + N_t \tag{3.4}$$

Here, x_{i-1} represents the output features from the previous layer of the network. "Conv" and "Norm" refer to 2D convolutional and normalization layers, with group normalization (GN) used for normalization. σ denotes a nonlinear activation function, with GELU [36] serving as the activation function. N_e and N_t correspond to the number of encoder and spatiotemporal learning modules, respectively.

The sub-band images obtained from each wavelet transformation layer serve as inputs for the convolutional layer. This convolutional layer's purpose is to acquire concise representations for channel feature compression. The features we obtain from the image are essential information about the images like edges, texture and structure. The compressed features, in turn, are utilized as inputs for the subsequent wavelet transformation to achieve frequency decomposition. Notably, the wavelet transform replaces the typical pooling operation, which is often used for downsampling. This replacement offers a unique advantage: the network can perform subsampling without any loss of information. The frequency and positional characteristics obtained through the DWT are particularly advantageous in preserving texture features when compared to convolutional layer downsampling with a stride of two. This feature preservation is essential in applications where maintaining the integrity of the original image is crucial, such as in high-resolution video processing or detailed spatial analysis.

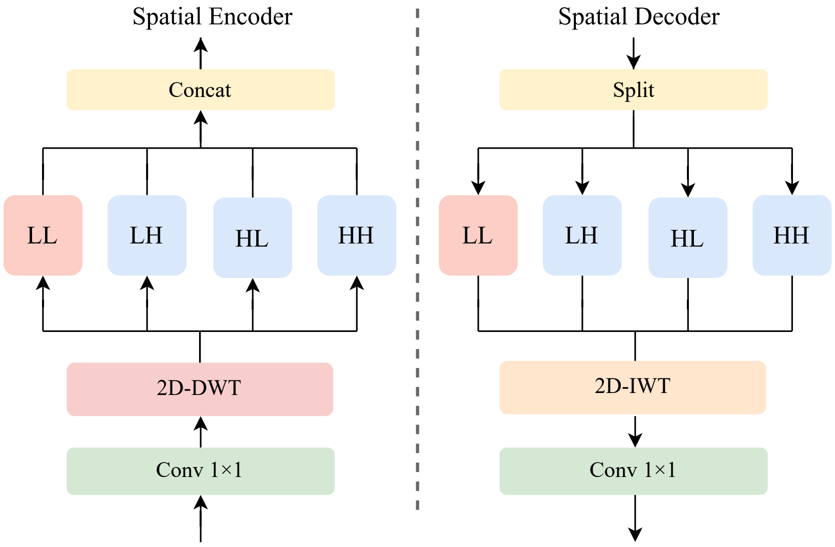

Figure 3. The overview of spatial encoder-decoder based on wavelet.

The proposed STU framework, with its multilevel wavelet transform encoder-decoder, offers a novel approach to video frame prediction and spatial feature extraction. This methodology has significant potential for various applications, including but not limited to, dynamic system monitoring, high-definition video processing, and detailed spatial analysis in scientific research. By combining the strengths of wavelet transforms with advanced neural network architectures, the STU framework sets a new standard in the field of video analysis and prediction.

3.2. Spatiotemporal attention unit

The spatiotemporal learning module accepts input potential features from the spatial encoder f and produces potential spatiotemporal representations for decoding by the spatial decoder f^{-1}. In STU, a spatiotemporal learning module grounded in pure convolutional networks is introduced to extract these potential spatiotemporal representations.

In this study, we adopt the MetaFormer [31] design principles, which commonly entail a token mixer and a channel mixer composition. The token mixer is responsible for extracting spatial features, while the channel mixer focuses on extracting temporal features. Specifically, as shown in Figure 4, the token mixer employs DW Conv with an 11×11 convolution kernel and a dilation rate of 3. The utilization of a larger convolution kernel facilitates a broader sensory field coverage, while the use of DW Conv aims to reduce the computational load of model parameters. Additionally, the DW Conv layers are interleaved with batch normalization and ReLU activation functions to enhance the nonlinearity and generalization capability of the model. The channel mixer consists of two convolutional layers with a 1×1 convolutional kernel, enabling efficient feature fusion across the temporal dimension. This design

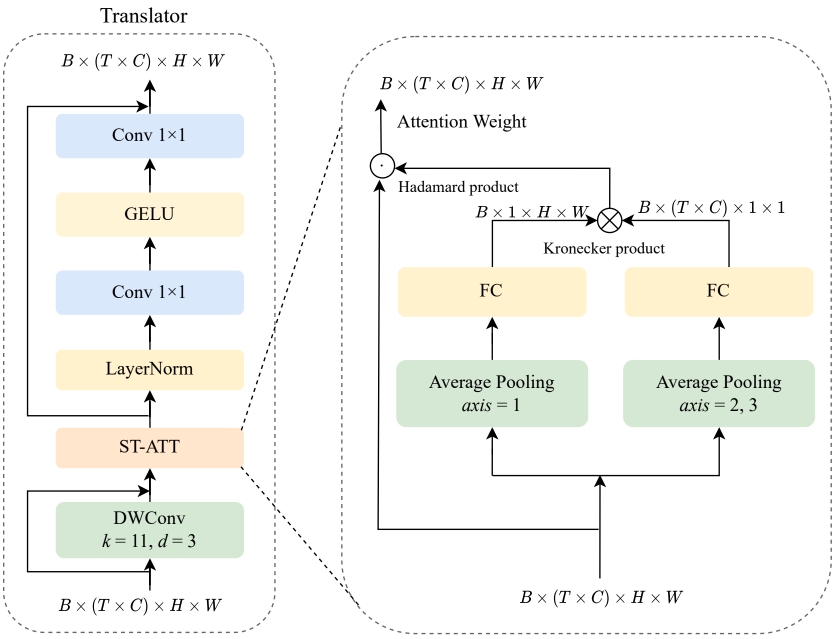
Translator
$B \times (T \times C) \times H \times W$
Conv 1×1
GELU
Conv 1×1
LayerNorm
ST-ATT
DWConv
$k = 11, d = 3$
$B \times (T \times C) \times H \times W$
$B \times (T \times C) \times H \times W$
Attention Weight
Hadamard product
$B \times 1 \times H \times W$
$B \times (T \times C) \times 1 \times 1$
Kronecker product
FC
FC
Average Pooling
$axis = 1$
Average Pooling
$axis = 2, 3$
$B \times (T \times C) \times H \times W$

Figure 4. The architecture of spatiotemporal translator module.

choice effectively captures the dynamic temporal patterns essential for accurate motion interpretation in video data.

Moreover, the architecture incorporates residual connections in both mixers, enhancing the gradient flow during backpropagation and preventing the vanishing gradient problem. This is particularly beneficial for training deeper models where information needs to be propagated over many layers.

The spatiotemporal learning module can be represented as depicted in the following equation:

$$z_i = \mathrm{Norm}\left(\mathrm{DWConv}_{k=11,d=3}\left(x_{i-1}\right)\right) \tag{3.5}$$

$$x_i = \mathrm{Conv}_{k=1}\left(\sigma\left(\mathrm{Conv}_{k=1}\left(z_i\right)\right)\right) + x_{i-1} \tag{3.6}$$

Power video surveillance scenes often feature multiple objects with complex motion in the video. This paper introduces a time-frequency-based attention mechanism to ensure both temporal and spatial continuity. The attention mechanism is decomposed into an SAU capturing spatial frequency within frames and a TAU handling relationships between frames. This bifurcated approach allows for more precise and adaptive attention allocation, significantly improving the model's ability to focus on relevant spatiotemporal features while disregarding noise and irrelevant information.

For the spatial frequency features extracted by the wavelet encoder, two steps are taken to generate attention vectors. First, global time channel pooling is applied, producing the spatial attention vector $\mathcal{A}_{\mathrm{SAU}}$ by compressing time channel information via a linear layer. In the second step, global pooling is again employed, but this time on the spatial axes (i.e., axes 2 and 3). This results in the spatial information being compressed and passed through a linear layer to generate the time channel attention vector $\mathcal{A}_{\mathrm{TAU}}$. The innovative use of global pooling across different axes enables the model to capture a comprehensive view of the spatial-temporal landscape within the video, allowing for more nuanced and detailed feature extraction.

Finally, the attention vectors from the two steps are processed through the Kronecker product to create the spatiotemporal attention matrix $\mathcal{A}$. This matrix plays a crucial role in modulating the feature representations, ensuring that the model's focus is dynamically adjusted to the most salient aspects of the input data. This is mathematically represented as follows:

$$\mathcal{A}_{\text{SAU}} = \text{FC}(\text{AvgPool}_{\text{axis}=1}(x_{i-1})) \tag{3.7}$$

$$\mathcal{A}_{\text{TAU}} = \text{FC}(\text{AvgPool}_{\text{axis}=2,3}(x_{i-1})) \tag{3.8}$$

$$\mathcal{A} = \mathcal{A}_{\text{SAU}} \otimes \mathcal{A}_{\text{TAU}} \tag{3.9}$$

In these equations, $\mathcal{A}_{\text{SAU}}$ is a dynamic spatial attention vector with dimensions $\mathbb{R}^{B\times 1\times H\times W}$, and $\mathcal{A}_{\text{TAU}}$ is a dynamic temporal attention vector with dimensions $\mathbb{R}^{B\times(T\times C)\times 1\times 1}$. The FC and AvgPool denote fully connected (FC) and average pooling layer, respectively. These components are responsible for creating the spatiotemporal attention matrix, which will be later combined with the token mixer's output through a Hadamard product, as expressed in the following formula:

$$x_i = x_{i-1} \odot \mathcal{A} \tag{3.10}$$

This comprehensive approach to spatiotemporal feature extraction and attention allocation significantly enhances the model's ability to analyze complex video data, making it highly suitable for advanced applications such as intelligent surveillance, autonomous vehicle navigation, and human-computer interaction systems.

4. Experiments

In this paper, we will show the results of experiments on publicly available video prediction benchmark datasets and power surveillance video prediction datasets. The experiments are implemented according to different dataset setups to evaluate the performance of the proposed model.

4.1. Experimental data

The evaluation of our model spans a diverse array of datasets, encompassing both synthetic and realistic scenes, to validate its efficacy and versatility. The datasets used in this study include:

- Moving MNIST [37]: A synthetic dataset of two digits moving within a grid. The digits bounce off the boundaries, and the model is required to learn the pattern of digit movement and reconstruct the prediction frames. The sample size of the used data can be seen in the Table 2. This dataset is widely used as a benchmark in standard spatiotemporal prediction learning tasks.
- PowerAction: A private dataset accumulated over time by the laboratory in the field of electric power. It is primarily based on electric power scenes such as substations. The dataset consists of a total of 21,124 sample images, derived from video sampling. It contains a variety of indoor and outdoor scenes, six types of personnel behaviors (climbing, climbing ladder, smoking, falling, crossing, moving), and scenarios with different numbers of people (single, multiple) to simulate a wide range of real-world situations. The main difference between the common power scenarios and complex power scenarios we deal with is that complex scenarios may combine more than one behaviors and have a different number of people. The sample distribution of the dataset is detailed in Table 1.

- TrafficBJ: This dataset comprises a comprehensive collection of traffic condition data for the road network within Beijing, sourced from Baidu Map. This dataset features two distinct channels: inflow and outflow of traffic. In alignment with methodologies adopted in prior research, this study has normalized the data to a range between 0 and 1, facilitating more straightforward analysis and comparison.

These datasets provide a robust platform for evaluating the performance of our proposed model. The Moving MNIST dataset allows us to test the model's ability to learn and predict simple, yet dynamic patterns, while the PowerAction dataset presents a more complex challenge, requiring the model to understand and predict human behaviors and interactions in various real-world power facility scenarios. The results of these experiments will demonstrate the model's capability to handle both synthetic and realistic video prediction tasks.

Table 1. Sample size of the PowerAction dataset.

Behavioral categories	Sample size (statistics)
climb high	3645
climbers	3201
cigarette smoking	3520
fall to the ground	3910
straddle	3587
mobility	3261

Table 2. Sample sizes of other used datasets.

Dataset	Sample size
Moving MNIST	20,000
TrafficBJ	20,961

4.2. Experimental details

Referring to the previous evaluation metrics [21], in this paper, we use MSE, structure similarity index measure (SSIM), and peak signal-to-noise ratio (PSNR) to assess the prediction quality. MSE and MAE assess the pixel-by-pixel error, SSIM measures the similarity of structural information within the spatial domain, and PSNR measures the similarity of structural information within the spatial domain and assesses the quality of the prediction. PSNR is the ratio of the maximum possible power of the signal to the power of the distorted noise.

In this study, we implement the proposed model using the PyTorch framework and conduct training on a single NVIDIA-A100 GPU. The model training employs a batch size of 16 video sequences. We utilize the AdamW optimizer with a learning rate of 0.001 and a weight decay of 0.05. The learning rate is fine-tuned using a combination of WarmUp and CosineAnneal techniques.

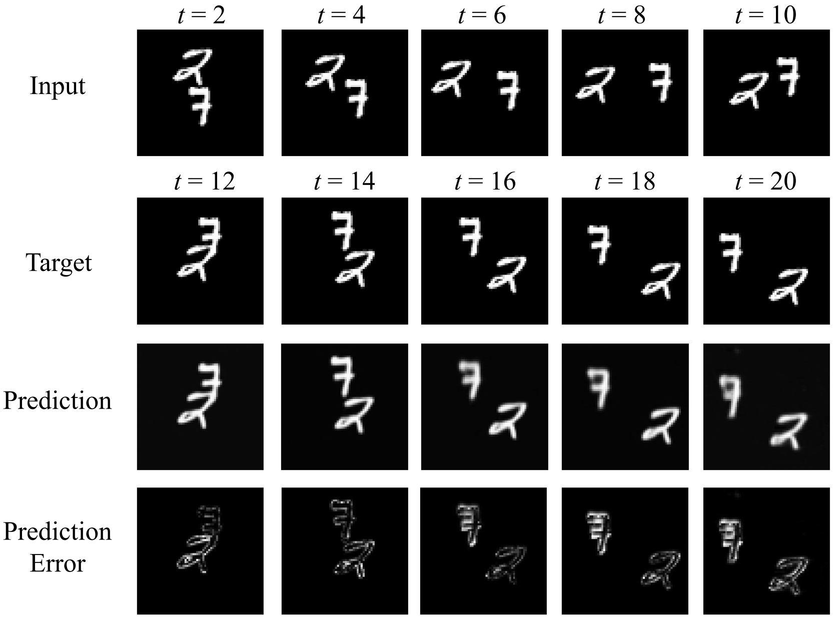

Figure 5. Moving MNIST experiment visualization results.

4.3. Experimental results

This study assesses the proposed model's performance against recent robust baseline models, including popular RNN-based models such as ConvLSTM [17], PredRNN [18], PredRNN++ [19], MIM [24], E3D-LSTM [20], PhyDNet [21], SimVP [32], and TAM [33]. The following quantitative analyses are conducted on the PowerAction datasets.

4.3.1. Moving MNIST dataset

We first conduct several ablation experiments on DWT, as detailed in Table 3. The number of output channels and AvgPool and FC layers prove the feasibility of our model, as shown in Tables 4 and 5. The former two are about the encoder-decoder and the last one is the about translator.

Table 3. Ablation experiment on DWT.

	MSE	SSIM
Using DWT	20.8	0.949
Without using DWT	22.9	0.931

In the study focusing on the Moving MNIST dataset, a detailed experimental performance comparison was conducted to evaluate various models, with the results presented in Table 6. This table provides a clear comparison between the proposed model in this research and several established models in terms of MSE and SSIM.

The ConvLSTM [17] showed an MSE of 103.3 and an SSIM of 0.707, indicating a baseline level of performance in this context. PhyDNet [21] significantly improved upon these metrics, achieving

Table 4. Ablation experiment on numbers of output channels.

	MSE	SSIM
Only LL	23.8	0.886
LL & LH	22.5	0.891
LL & LH & HL	21.6	0.915
LL & LH & HL & HH	20.8	0.949

Table 5. Ablation experiment on translator.

	MSE	SSIM
Without AvgPool (axis = 1)	22.3	0.912
Without AvgPool (axis = 2,3)	23.4	0.897
Without FC	25.8	0.862
Ours	20.8	0.949

Table 6. Experimental performance comparison of moving MNIST dataset.

Method	**MSE**	**SSIM**
ConvLSTM [17]	103.3	0.707
PhyDNet [21]	24.4	0.947
PredRNN [18]	64.1	0.870
MIM [24]	44.2	0.910
PredRNN++ [19]	46.5	0.898
E3D-LSTM [20]	41.3	0.910
SimVP [32]	26.8	0.912
TAM [33]	23.4	0.934
Ours	**20.8**	**0.949**

an MSE of 24.4 and an SSIM of 0.947. This demonstrated a substantial enhancement in both error reduction and structural similarity. PredRNN [18] delivered an MSE of 64.1 and an SSIM of 0.870. Its successor, PredRNN++ [19], showed comparable performance with an MSE of 46.5 and an SSIM of 0.898. Another notable model, MIM [24], exhibited an MSE of 44.2 and an SSIM of 0.910, indicating its competitive performance in capturing the dynamics of the Moving MNIST dataset. The E3D-LSTM [20] demonstrated its effectiveness with an MSE of 41.3 and an SSIM of 0.910. From the above results, we can find that since RNN is based on sequence structure, RNN-based models are generally inferior to CNN-based models in image processing. CNN-based models have operations like convolution and pooling, which are effectively good at extracting features from images and thus process images well. However, the model proposed in this paper achieved the most notable performance, outstripping all others with the lowest MSE of 20.8 and the highest SSIM of 0.949. These results clearly illustrate that the proposed model not only substantially reduces the prediction error but also significantly enhances the structural similarity, thus leading to a considerable performance gain over the baseline and other advanced models.

Furthermore, the qualitative visualization of the prediction results, as illustrated in Figure 5, provides additional insights into the performance of the proposed model. This visualization is crucial as it offers a tangible representation of how well the model can predict the motion in the Moving MNIST dataset. In summary, the experimental results highlight the superiority of the proposed model in handling the Moving MNIST dataset. Its ability to achieve the lowest MSE suggests a high degree of accuracy in predicting the movement of digits, which is a central aspect of the dataset. Simultaneously, the highest SSIM indicates that the proposed model is exceptionally proficient in maintaining the structural integrity of the digits during prediction, which is a critical measure of performance in video prediction tasks. These results collectively underscore the effectiveness of the proposed model in tackling the complex dynamics and nuances of the Moving MNIST dataset, setting a new benchmark in this area of research.

4.3.2. TrafficBJ

In the analysis of the TrafficBJ dataset, as detailed in Table 7, the experimental results highlight the efficacy of the proposed model in this study. The visual comparison of prediction results, as shown in Figure 6, reveals the model's ability to accurately predict future events despite notable differences between input and predicted frames. The discrepancies are mainly around the middle of the frames, but the overall trends closely match the real frames. This demonstrates the model's strong spatial-temporal learning capabilities, especially in real-world scenarios, underscoring its practical applicability in urban traffic analysis and forecasting.

In the performance comparison, as seen in the table, various methods are evaluated using the MSE and MAE metrics. The ConvLSTM [17], PhyDNet [21], PredRNN [18], MIM [24], PredRNN++ [19], and E3D-LSTM [20] methods present varying levels of performance. However, our proposed method significantly outperforms the others with the lowest MSE of 35.6 and MAE of 15.8. This superior performance not only reflects the model's precision in predicting traffic conditions but also highlights its potential as a robust tool for traffic management and urban planning, offering valuable insights for future advancements in the field.

Table 7. Experimental performance comparison of TrafficBJ dataset.

Method	MSE	MAE
ConvLSTM [17]	48.5	17.7
PhyDNet [21]	41.9	16.2
PredRNN [18]	46.4	17.1
MIM [24]	42.9	16.6
PredRNN++ [19]	44.8	16.9
E3D-LSTM [20]	43.2	16.9
SimVP [32]	37.8	16.5
TAM [33]	36.7	16.2
Ours	**35.6**	**15.8**

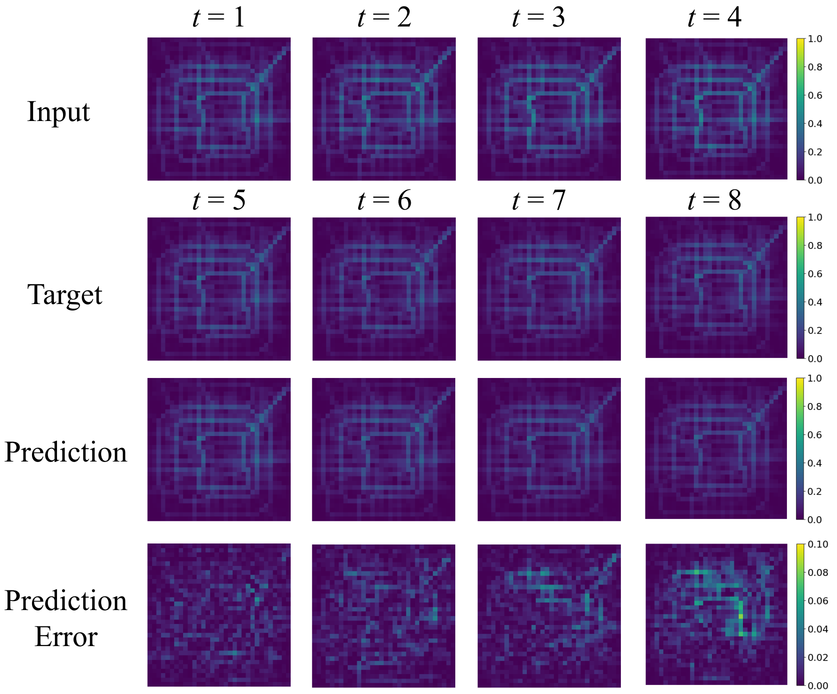

Figure 6. TrafficBJ experiment visualization results.

4.3.3. PowerAction dataset

In the study involving the PowerAction dataset, the experimental outcomes are meticulously detailed in Table 8, offering a comparative analysis of various predictive models. This dataset, known for its complex motion scenarios related to power-related locations, presents a significant challenge in predictive modeling. The research employs a pretrained Yolov5s target detection model, as described in Yolov5s [34], to assess the quality of the model's predictions and to detect potential rule violations. With a confidence threshold set at 0.7, the detection of frames serves as an indirect indicator of the perceptual quality of the predicted frames. The visual representations in Figure 7 demonstrate that the detection model is capable of producing high-quality output frames, reflecting the favorable visual quality of the predictions.

The experimental results, as shown in Table 8, include the performance of various models evaluated using the metrics SSIM and PSNR. The ConvLSTM [17], presents an SSIM of 0.731 and a PSNR of 23.8. On the other hand, PhyDNet [21] achieves an SSIM of 0.814 and a PSNR of 25.8, indicating a notable performance improvement. The PredRNN model [18] further advances these metrics, recording an SSIM of 0.849 and a PSNR of 27.6. In contrast, the MIM model [24] shows an SSIM of 0.761 and a PSNR of 24.9. Notably, the PredRNN++ model [19] delivers an SSIM of 0.874 and a PSNR of 28.3, while the E3D-LSTM model [20] leads with an SSIM of 0.889 and a PSNR of 29.2. The PowerAction dataset is a more complex dataset in comparison to the above two datasets. We can find that in this dataset, even the best RNN-based model is still inferior to the CNN-based models. The reason is generally the same as what we have mentioned in the Moving MNIST dataset: The RNN-based model has sequence structure and is not as good at image processing as the CNN-based model.

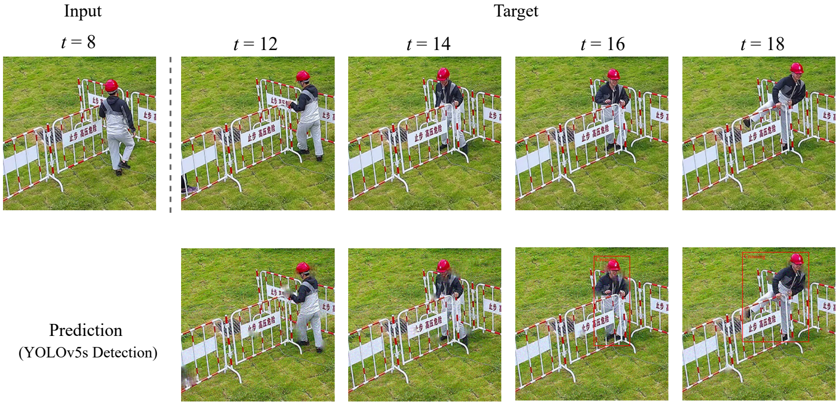

Figure 7. PowerAction experiment visualization results.

Among all, the model proposed in this research outshines the rest, achieving the highest SSIM of 0.921 and an impressive PSNR of 34.3. These results highlight the exceptional capability of the proposed model in maintaining structural integrity and enhancing the clarity and quality of predictions, making it particularly suited for handling the complexities of the PowerAction dataset. The combination of quantitative metrics and qualitative visualizations in this study underscores the significant advancements made in predictive modeling for complex motion scenarios.

Table 8. Experimental performance comparison of PowerAction dataset.

Method	SSIM	PSNR
ConvLSTM [17]	0.731	23.8
PhyDNet [21]	0.814	25.8
PredRNN [18]	0.849	27.6
MIM [24]	0.761	24.9
PredRNN++ [19]	0.874	28.3
E3D-LSTM [20]	0.889	29.2
SimVP [32]	0.901	31.2
TAM [33]	0.912	32.1
Ours	**0.921**	**34.3**

5. Conclusions

Given the suboptimal state of power violation identification, there is an imminent demand for early detection of hazardous behaviors in practical scenarios. In response, this study introduces a novel spatiotemporal prediction framework employing full convolution, thereby replacing high-latency cyclic structures. This transformation enables parallel reasoning of future video frames and significantly enhances model inference speed. Furthermore, we introduce a multilevel frequency domain codec, tailored to the dynamic electric power environment, for improved learning of spatial dependencies and temporal dynamics. This is accomplished by establishing time-frequency relationships within and between consecutive video frames. For power video surveillance scenarios, which frequently involve multiple objects, we propose the incorporation of a token mixer with an extensive sensory field. To ensure temporal and spatial continuity, we decompose attention into SAUs within frames and TAUs between frames. This approach surpasses existing methods in terms of accuracy and speed across public datasets, standard spatiotemporal prediction tasks, and cross-dataset generalization. Experimental results on PowerAction, a behavioral dataset within the electric power domain, demonstrate that STU exhibits promising practical applications. This paper establishes a robust baseline for power video prediction research and introduces innovative perspectives for realizing power violation recognition.

Future research in electricity rule violation detection should focus on optimizing computational efficiency, adapting to dynamic environments, understanding complex motion scenes, enhancing model robustness, integrating with existing systems, ensuring scalability and ease of deployment, and addressing data privacy concerns. Researchers should aim to reduce computational load using lightweight model architectures, knowledge distillation, model pruning, and hardware accelerators like GPUs and TPUs. Adapting to real-world environments requires online and reinforcement learning for continuous updates and transfer learning for different grid conditions. Advanced techniques combining CNNs and RNNs are crucial for understanding complex motions and scenes, while synthetic data generation can diversify training datasets. Ensuring robustness involves extensive testing and developing anomaly detection mechanisms. Integration with existing systems necessitates standardized APIs and communication protocols, and collaboration with industry partners. Scalability and deployment can be achieved by designing architectures suitable for edge devices and cloud platforms and creating automated deployment pipelines. Addressing data privacy involves robust encryption and secure data transmission to comply with regulations. Focusing on these areas will enhance the effectiveness and reliability of detection systems, support safe power grid operations, and promote industry advancements in intelligence and security, enabling effective performance in complex environments.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by the State Grid Zhejiang Electric Power Co., Ltd. Science and Technology Project: Edge Cloud Live Platform for Characterization Defect Detection of Substation Equipment Based on Machine Vision (Grant No. 5211DS220003).

Conflict of interest

The authors declare there is no conflicts of interest.

References

1. J. Hu, B. Guo, W. Yan, J. Lin, C. Li, Y. Yan, A classification model of power operation inspection defect texts based on graph convolutional network, *Front. Energy Res.*, **10** (2022), 1028607. https://doi.org/10.3389/fenrg.2022.1028607

2. Y. Yan, Y. Han, D. Qi, J. Lin, Z. Yang, L. Jin, Multi-label image recognition for electric power equipment inspection based on multi-scale dynamic graph convolution network, *Energy Rep.*, **9** (2023), 1928–1937. https://doi.org/10.1016/j.egyr.2023.04.152

3. Z. Chang, X. Zhang, S. Wang, S. Ma, W. Gao, STRPM: A spatiotemporal residual predictive model for high-resolution video prediction, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 13946–13955.

4. R. Villegas, J. Yang, S. Hong, X. Lin, H. Lee, Decomposing motion and content for natural video sequence prediction, preprint, arXiv:1706.08033.

5. Y. Zhang, Y. Yan, G. Feng, Feature compensation network based on non-uniform quantization of channels for digital image global manipulation forensics, *Signal Process. Image Commun.*, **107** (2022), 116795. https://doi.org/10.1016/j.image.2022.116795

6. T. Tao, K. Long, T. Yang, S. Liu, Y. Yang, X. Guo, et al., Quantitative assessment on fatigue damage induced by wake effect and yaw misalignment for floating offshore wind turbines, *Ocean Eng.*, **288** (2023), 116004. https://doi.org/10.1016/j.oceaneng.2023.116004

7. T. Tao, Y. Liu, Y. Qiao, L. Gao, J. Lu, C. Zhang, et al., Wind turbine blade icing diagnosis using hybrid features and Stacked-XGBoost algorithm, *Renewable Energy*, **180** (2021), 1004–1013. https://doi.org/10.1016/j.renene.2021.09.008

8. D. Ha, J. Schmidhuber, Recurrent world models facilitate policy evolution, *Advances in Neural Information Processing Systems*, **31** (2018).

9. S. Gidaris, P. Singh, N. Komodakis, Unsupervised representation learning by predicting image rotations, preprint, arXiv:1803.07728.

10. D. Song, Y. Yang, S. Zheng, X. Deng, J. Yang, M. Su, et al., New perspectives on maximum wind energy extraction of variable-speed wind turbines using previewed wind speeds, *Energy Convers. Manage.*, **206** (2020), 112496. https://doi.org/10.1016/j.enconman.2020.112496

11. A. Gupta, S. Tian, Y. Zhang, J. Wu, R. Martín-Martín, F. Li, Maskvit: Masked visual pre-training for video prediction, preprint, arXiv:2206.11894.

12. A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid, ViViT: A video vision transformer, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2021), 6836–6846.

13. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 16000–16009.

14. C. Tan, S. Li, Z. Gao, W. Guan, Z. Wang, Z. Liu, et al., OpenSTL: A comprehensive benchmark of spatio-temporal predictive learning, preprint, arXiv:2306.11249.

15. S. Jenni, G. Meishvili, P. Favaro, Video representation learning by recognizing temporal transformations, in *European Conference on Computer Vision*, (2020), 425–442.

16. L. Castrejon, N. Ballas, A. Courville, Improved conditional VRNNs for video prediction, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 7608–7617.

17. X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, W. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, in *Advances in Neural Information Processing Systems*, **28** (2015).

18. Y. Wang, M. Long, J. Wang, Z. Gao, P. S. Yu, PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMS, in *Advances in Neural Information Processing Systems*, **30** (2017).

19. Y. Wang, Z. Gao, M. Long, J. Wang, S. Y. Philip, PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning, in *International Conference on Machine Learning*, (2018), 5123–5132.

20. Y. Wang, L. Jiang, M. H. Yang, L. Li, M. Long, F. Li, Eidetic 3D LSTM: A model for video prediction and beyond, in *International Conference on Learning Representations*, 2018.

21. V. Le Guen, N. Thome, Disentangling physical dynamics from unknown factors for unsupervised video prediction, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 11474–11484.

22. F. Ebert, C. Finn, A. X. Lee, S. Levine, Self-supervised visual planning with temporal skip connections, *CoRL*, **12** (2017), 16.

23. S. Liu, Z. Lin, Y. Zhao, Y. Liu, Y. Ding, B. Zhang, et al., Robust system separation strategy considering online wide-area coherency identification and uncertainties of renewable energy sources, *IEEE Trans. Power Syst.*, **35** (2020), 3574–3587. https://doi.org/10.1109/TPWRS.2020.2971966

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

25. M. Chen, Y. Wang, Y. Dai, Y. Yan, D. Qi, Small and strong: Power line segmentation network in real time based on self-supervised learning, *Proc. CSEE*, **42** (2022), 1365–1375.

26. X. Wang, L. Luo, L. Tang, Z. Yang, Automatic representation and detection of fault bearings in in-wheel motors under variable load conditions, *Adv. Eng. Inf.*, **49** (2021), 101321. https://doi.org/10.1016/j.aei.2021.101321

27. I. Xu, Y. Wang, M. Long, J. Wang, PredCNN: Predictive learning with cascade convolutions, in *IJCAI*, (2018), 2940–2947.

28. J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7132–7141.

29. T. Yao, Y. Pan, Y. Li, C. Ngo, T. Mei, Wave-ViT: Unifying wavelet and transformers for visual representation learning, in *European Conference on Computer Vision*, (2022), 328–345.

30. W. Lotter, G. Kreiman, D. Cox, Deep predictive coding networks for video prediction and unsupervised learning, preprint, arXiv:1605.08104.

31. W. Yu, M. Luo, P. Zhou, C. Si, Y. Zhou, X. Wang, et al., Metaformer is actually what you need for vision, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 10819–10829.

32. Z. Gao, C. Tan, L. Wu, S. Z. Li, SimVP: Simpler yet better video prediction, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 3170–3180.

33. X. Nie, X. Chen, H. Jin, Z. Zhu, Y. Yan, D. Qi, Triplet attention transformer for spatiotemporal predictive learning, in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, (2024), 7036–7045.

34. G. Jocher, A. Chaurasia, A. Stoken, J. Borovec, Y. Kwon, K. Michael, et al., ultralytics/yolov5: v7.0-YOLOv5 sota realtime instance segmentation, *Zenodo*, 2022.

35. C. Tan, Z. Gao, S. Li, S. Z. Li, SimVP: Towards simple yet powerful spatiotemporal predictive learning, preprint, arXiv:2211.12509.

36. D. Hendrycks, K. Gimpel, Gaussian error linear units (GELUs), preprint, arXiv:1606.08415.

37. J. Lee, J. Lee, S. Lee, S. Yoon, Mutual suppression network for video prediction using disentangled features, preprint, arXiv:1804.04810.

AIMS Press