Wavelet-Driven Spatiotemporal Predictive Learning: Bridging Frequency and Time Variations

Xuesong Nie[1], Yunfeng Yan[1*], Siyuan Li[1,2], Cheng Tan[1,2], Xi Chen[3], Haoyuan Jin[1], Zhihang Zhu[1], Stan Z. Li[2], Donglian Qi[1]

[1]Zhejiang University, Zhejiang, China
[2]School of Engineering, Westlake University, Zhejiang, China
[3]Department of Computer Science, The University of Hong Kong, Hong Kong, China
xuesongnie@zju.edu.cn

Abstract

Spatiotemporal predictive learning is a paradigm that empowers models to learn spatial and temporal patterns by predicting future frames from past frames in an unsupervised manner. This method typically uses recurrent units to capture long-term dependencies, but these units often come with high computational costs and limited performance in real-world scenes. This paper presents an innovative Wavelet-based SpatioTemporal (WaST) framework, which extracts and adaptively controls both low and high-frequency components at image and feature levels via 3D discrete wavelet transform for faster processing while maintaining high-quality predictions. We propose a Time-Frequency Aware Translator uniquely crafted to efficiently learn short- and long-range spatiotemporal information by individually modeling spatial frequency and temporal variations. Meanwhile, we design a wavelet-domain High-Frequency Focal Loss that effectively supervises high-frequency variations. Extensive experiments across various real-world scenarios, such as driving scene prediction, traffic flow prediction, human motion capture, and weather forecasting, demonstrate that our proposed WaST achieves state-of-the-art performance over various spatiotemporal prediction methods. Our code is available at https://github.com/xuesongnie/WaST.

Introduction

Spatiotemporal predictive learning has recently seen significant progress. Central to data-driven spatiotemporal predictive learning is the generation of future frames based on historical frames, with applications including traffic flow prediction (Fang et al. 2019), weather forecasting (Reichstein et al. 2019), physical scene comprehension (Xu et al. 2019), early activity recognition (Wang et al. 2018b), and vision-based predictive control (Gupta et al. 2022). Leveraging these massive unlabeled data, these models can self-supervised uncover complex spatial and temporal dependencies, forecasting future events from past data. With the abundance of spatiotemporal data, unsupervised pretraining strategies using these techniques show promise for enhancing universal representation learning across various visual tasks (Tan et al. 2023b; Li et al. 2023).

*Corresponding author.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Spatiotemporal predictive learning is commonly modeled with two methods: *recurrent-based* and *recurrent-free* frameworks shown in Figure 1. Many mainstream recurrent-based methods (Chang et al. 2021; Wang et al. 2022) suggested leveraging models with stacked recurrent units to capture the temporal dependencies. Inspired by the success of long short-term memory (LSTM) networks (Hochreiter et al. 1997) in sequence modeling, ConvLSTM (Shi et al. 2015), PredRNN (Wang et al. 2017), PredRNN++ (Wang et al. 2018a), and MIM (Wang et al. 2019) propose various variants to improve the performance of vanilla LSTM, called MetaLSTM, such as ConvLSTM, ST-LSTM, CausalLSTM, and MIM-LSTM. Thus we abstract the general framework of recurrent-based models shown in Figure 1, which consists of two main parts: (i) various LSTM variants called MetaLSTM. (ii) the information flow modes between different recurrent units. Despite the advantages of the recurrent-based framework in long-term prediction, the heavy computational effort limits its further application. Recently, recurrent-free methods (Tan et al. 2022, 2023a) with the advantage of parallelization has been proposed for spatiotemporal learning. As shown in Figure 1(a), we demonstrated the recurrent-free framework representing various variants of SimVP (Gao et al. 2022), which also consists of two main parts: (i) spatial 2D Encoder-Decoder. (ii) latent feature Spatio-Temporal Translator. Despite greater computational efficiency, the above methods have performance gaps in real-world scenarios due to single-scale architecture, 2D operations, and irrobust spatiotemporal translators.

In this work, we present a novel Wavelet-based SpatioTemporal (WaST) prediction scheme to tackle existing performance gaps. Our solution relies on multi-level 3D Discrete Wavelet Transform (3D-DWT), which decomposes data into low- and high-frequency wavelet subbands. We introduce 3D-Wavelet Embedding and Reconstruction modules to embed the wavelet prior and perform detail-oriented reconstruction respectively. To learn intricate frequency features and temporal dynamics, we propose a Time-Frequency Aware Translator (TF-Aware Translator) realized via frequency-mixer and temporal-mixer to extract and adaptively control both low- and high-frequency components in multi-level wavelet space. We argue that traditional mean square error loss makes it hard to focus on de-

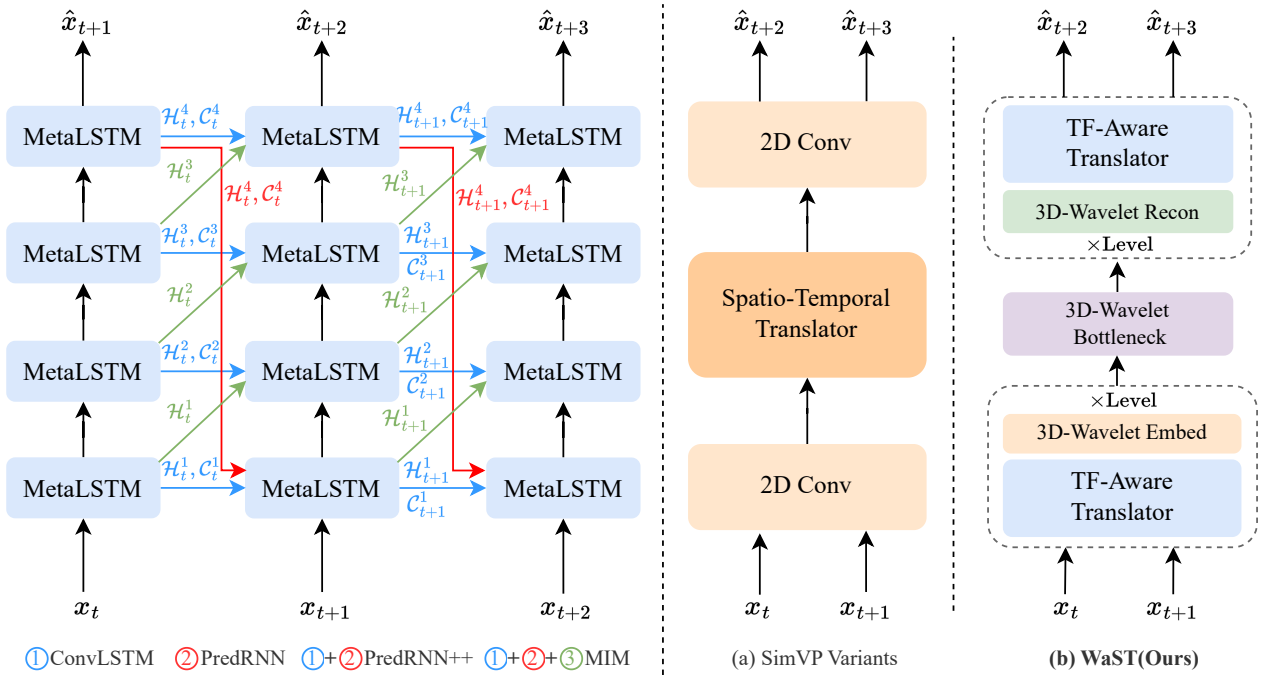


Figure 1: Two typical spatiotemporal predictive learning frameworks. The recurrent-based methods extract spatiotemporal dependencies through recurrent units, notably MetaLSTM, and the information flows between these units. The recurrent-free such as (a) SimVP variants extract spatiotemporal features through elaborate Spatio-Temporal Translators in latent space. In contrast, (b) our proposed WaST extracts and adaptively controls both low and high-frequency components via time-frequency aware translators in multi-level wavelet space.

tailed features, hence we introduce a wavelet-domain High-Frequency Focal Loss (HFFL) that also captures rapid alterations between consecutive and static frames in the wavelet domain. The above innovations have enabled our proposed method to achieve state-of-the-art performance in diverse real-world dynamic scenarios. We outline our key contributions as follows:

- A pioneering Wavelet-based SpatioTemporal (WaST) predictive framework is proposed that enhances computational efficiency while ensuring high-quality predictions through high-frequency components.
- A Time-Frequency Aware Translator (TF-Aware Translator) is designed to efficiently learn real-world spatiotemporal dependencies by separately modeling spatial frequency and temporal variations.
- We propose a wavelet-domain High-Frequency Focal Loss (HFFL) to supervise high-frequency variations. Moreover, WaST provides state-of-the-art performance on various scenario datasets while being an order of magnitude faster than recurrent-based methods.

Related Work

Spatiotemporal Predictive Learning

The advances in recurrent-based models have significantly deepened our insight into spatiotemporal predictive learning. The pioneering work ConvLSTM (Shi et al. 2015) integrates convolutional networks with LSTM. PredRNN (Wang et al. 2017) and PredRNN++ (Wang et al. 2018a) utilized spatiotemporal LSTM (ST-LSTM) and gradient highway units to capture temporal dependencies while mitigating the gradient vanishing. MIM (Wang et al. 2019) using differential information between hidden states for better non-stationarity handling. E3D-LSTM (Wang et al. 2018b) incorporating 3D convolutions into LSTM. PhyDNet (Guen et al. 2020) disentangled PDE dynamics from unknown complementary information with a recurrent physical unit. MAU (Chang et al. 2021) designs a motion-aware unit to capture motion information. PredRNNv2 (Wang et al. 2022) employed a curriculum learning strategy and memory decoupling loss for enhanced performance.

In the recent past, recurrent-free models such as TAT (Nie et al. 2023), SimVP (Gao et al. 2022), TAU (Tan et al. 2023a), and DMVFN (Hu et al. 2023) have made advancements with triplet attention, inception module, temporal attention, and dynamic multi-scale voxel flow. Although these recurrent-free models have been developed, most works fail to consider the time-frequency dependence in real-world dynamic scenarios, leading to the missing of prediction details. We address these issues with the 3D wavelet framework and time-frequency aware translator.

Wavelet Transform in Computer Vision

Wavelet transform, essential for time-frequency analysis, bolsters performance in CNN-based visual tasks. (Oyallon et al. 2017) incorporates wavelet scattering networks into

ResNet (He et al. 2016). (Bae, et al. 2017) propose a wavelet transform to simplify topological structures of input or label manifolds for image restoration. DWSR (Guo et al. 2017) utilized low-resolution wavelet subbands for image super-resolution. WaveCNets (Li et al. 2020) and MWCNN (Liu et al. 2018) replaced traditional convolution operations with 2D wavelet transforms. Wave-ViT (Yao et al. 2022) incorporated 2D wavelet transform for better self-attention learning. Despite these advancements, the potential of wavelet transforms in spatiotemporal prediction remains under-explored. Therefore, we present a novel 3D wavelet-based scheme for better spatiotemporal learning.

Preliminaries

Problem Definition

The spatiotemporal predictive learning aims to model given past frames X_{in} to predict future frames $\widehat{X}_{\text{out}}$. We represent the spatiotemporal sequences as a four-dimensional tensor, *i.e.*, $X_{\text{in}}^{t:T} \in \mathbb{R}^{C\times T\times H\times W}$ and $X_{\text{out}}^{T+1:T+T'} \in \mathbb{R}^{C\times T'\times H\times W}$, where C, T, H and W denote channel, temporal or frames, height and width, respectively. The model with learnable parameters θ learns a mapping $\mathcal{F}_\theta : X_{\text{in}}^{t:T} \mapsto X_{\text{out}}^{T+1:T+T'}$ 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 predicted future frames and the ground-truth frames, the optimal parameters $\theta^\star$ are:

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

where $\mathcal{L}$ denotes loss function. In this paper, we take video prediction as a typical experimental domain, where the observed data are RGB images with three channels. In other domains, the observed data are multi-channel tensors.

Wavelet Transform

Wavelet Transform is a traditional technique that can separate low-frequency approximation and high-frequency details from the original data. In general, the wavelet transform process input signal involves two types of operations: Discrete Wavelet Transform (DWT) and Inverse Discrete Wavelet Transform (IDWT).

Opting for the Haar wavelet due to its simplicity, we utilize $\mathcal{F}_L = \frac{1}{\sqrt{2}}[\ 1 \quad 1\ ]$ and $\mathcal{F}_H = \frac{1}{\sqrt{2}}[\ -1 \quad 1\ ]$ as 3D-DWT low-pass and high-pass filters, respectively. Employing $\mathcal{F}_L$ and $\mathcal{F}_H$ to construct eight kernels with stride 2, $\mathcal{F}_{LLL}, \mathcal{F}_{LLH}, \mathcal{F}_{LHL}, \mathcal{F}_{LHH}, \mathcal{F}_{HLL}, \mathcal{F}_{HLH}, \mathcal{F}_{HHL}$ and $\mathcal{F}_{HHH}$. This framework enables the decomposition of input spatiotemporal sequences into eight downsampled subbands $X_{LLL}, X_{LLH}, X_{LHL}, X_{LHH}, X_{HLL}, X_{HLH}, X_{HHL}$ and X_{HHH}. Multi-level 3D-DWT further refines these subband components in a recurrent manner. To introduce wavelet technology into spatiotemporal predictive learning, we propose a 3D wavelet framework in Sec. and wavelet-domain High-Frequency Focal Loss (HFFL) in Sec. . Both approaches have been proven to enhance the predictive performance of the model.

Proposed Method

Overview

Take 1-level wavelet transform as an example, we present the overview architecture in Figure 2. The *Time-Frequency Aware Translator* (TF-Aware Translator) is implemented to learn spatial frequency and temporal variations in multi-level wavelet space. The *3D-Wavelet Embedding and Reconstruction* modules leverage wavelet's inherent multiscale architecture to embed and reconstruct intricate spatiotemporal features. The *Wavelet Bottleneck* block focuses on intermediate low-frequency representations while preserving the high-frequency details.

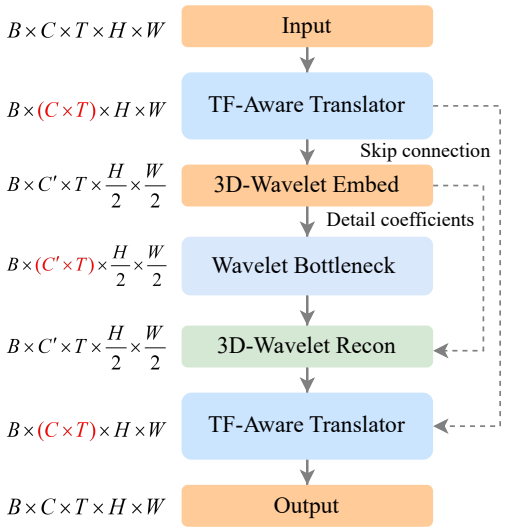


Figure 2: The overview architecture of our proposed WaST using the 1-level discrete wavelet transform.

Time-Frequency Aware Translator

The Time-Frequency Aware Translator (TF-Aware Translator) separates the modeling of intra-frame frequency and inter-frame temporal dynamic through Frequency-Mixer (FM) and Temporal-Mixer (TM), as shown in Figure 3. It reshapes input sequences $X \in \mathbb{R}^{B\times C\times T\times H\times W}$ to combine channel and temporal dimensions as $B\times(C\times T)\times H\times W$. With the two components mentioned above, the translator block can be formulated as follows:

$$X'_l = X_l + \text{diag}\left(\lambda_{l,1},\ldots,\lambda_{l,d}\right)\times \text{FM}\left(\eta\left(X_l\right)\right), \quad (2)$$

$$X_{l+1} = X'_l + \text{diag}\left(\lambda'_{l,1},\ldots,\lambda'_{l,d}\right)\times \text{TM}\left(\eta\left(X'_l\right)\right), \quad (3)$$

where $\lambda_{l,i}$ and $\lambda'_{l,i}$ are learnable parameters, and η is pre-normalization.

Frequency-Mixer. As illustrated at the top of Figure 3, the frequency-mixer is an ensemble of two key elements: the frequency feature extractor and frequency attention. Inspired by the recent metaformer modules (Ding et al. 2022; Liu et al. 2023; Li et al. 2022), we use parallel asymmetric convolution to fit oversize convolution kernels ($e.g.$, 51×51) to model low-frequency features and parallel small convolution kernels ($e.g.$, 5×5) to model high-frequency features.

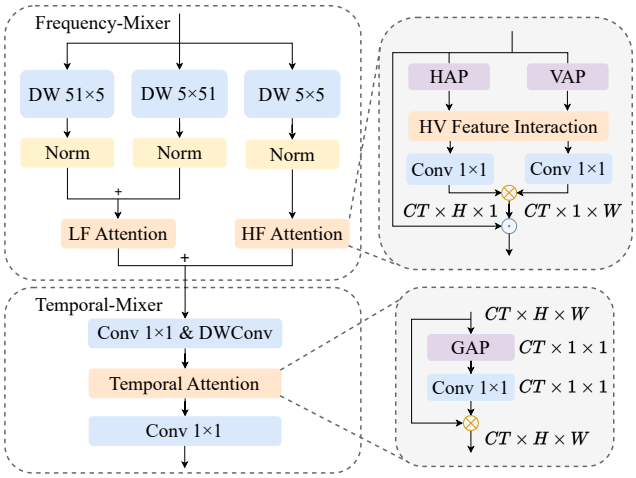


Figure 3: Illustration of Time-Frequency Aware Translator, which models intra-frame frequency and inter-frame temporal dynamic through Frequency-Mixer and Temporal-Mixer.

To enhance the representation of horizontal and vertical details in wavelets, we have proposed a frequency attention mechanism. To adaptively enhance wavelet features at different frequencies (*e.g.*, horizontal and vertical detail coefficients), the frequency attention is partitioned into horizontal and vertical components, refined by the squeeze-and-excitation paradigm (Hu et al. 2018), and interaction is promoted via weight-sharing 1×1 convolution, which can be formally expressed as:

$$\mathcal{A}_{\mathrm{H}} = W_{1\times1} * (W^{\star}_{1\times1} * \mathrm{HAP}(X)), \tag{4}$$

$$\mathcal{A}_{\mathrm{V}} = W_{1\times1} * (W^{\star}_{1\times1} * \mathrm{VAP}(X)), \tag{5}$$

$$X' = \mathcal{A}_{\mathrm{H}} \otimes \mathcal{A}_{\mathrm{V}} \odot X, \tag{6}$$

where $*$ is the convolution operation. HAP and VAP represent horizontal and vertical Average Pooling (AP). $\mathcal{A}_{\mathrm{H}} \in \mathbb{R}^{B\times(C\times T)\times H\times 1}$ and $\mathcal{A}_{\mathrm{V}} \in \mathbb{R}^{B\times(C\times T)\times 1\times W}$ denote the horizontal and vertical attention. $W_{1\times1}$ and $W^{\star}_{1\times1}$ both signify 1×1 convolutions, where $W^{\star}_{1\times1}$ is weight-sharing for horizontal and vertical attention interaction. The Kronecker and Hadamard products are represented by $\otimes$ and $\odot$.

Temporal-Mixer. As illustrated at the bottom of Figure 3, the temporal-mixer models long-term dependency by incorporating a feed-forward network (FFN) and temporal attention. The FFN contains two 1×1 convolutions and a depth-wise convolution (DWConv). The temporal attention models both channels and temporal dimensions for inter-frame dynamics in a squeeze-and-excitation (Hu et al. 2018) manner, it can be defined as:

$$\mathcal{A}_{\mathrm{T}} = W_{1\times1} * \mathrm{GAP}(X), \tag{7}$$

$$X' = \mathcal{A}_{\mathrm{T}} \otimes X, \tag{8}$$

where $\mathcal{A}_{\mathrm{T}} \in \mathbb{R}^{B\times(C\times T)\times1\times1}$ indicate the temporal attention, and GAP denotes global average pooling.

3D-Wavelet Embedding and Reconstruction

The 3D-Wavelet Embedding module integrates wavelet priors into the feature maps, while the Reconstruction module restores details from high-frequency coefficients. The

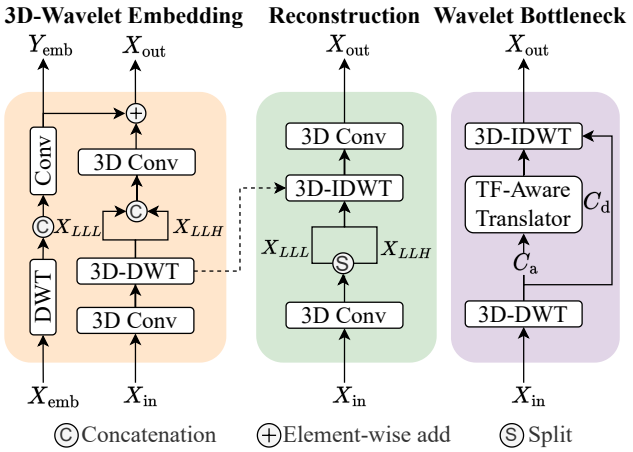


Figure 4: The detailed structure of the 3D-Wavelet Embedding and Reconstruction, and Wavelet Bottleneck modules.

Wavelet Bottleneck block emphasizes low-frequency representations while retaining the high-frequency details, as illustrated in Figure 4.

3D-Wavelet Embedding. We employ wavelet transform inherent properties and learnable 3D convolutions for better downsampling. Specifically, we designed two parallel branches, where one branch embeds image features X_{emb} into output features X_{out} via 3D-DWT and 3D Conv, while the other branch accomplishes spatial downsampling of the input feature maps X_{in}. During downsampling, we concatenate X_{LLL} and X_{LLH}, feeding remaining coefficients C_{detail} (*e.g.*, X_{LHL}, X_{LHH}, X_{HLL}, X_{HLH}, X_{HHL}, and X_{HHH}) into the Reconstruction module for detail reconstruction.

3D-Wavelet Reconstruction. During upsampling, we use both spatial low-frequency features X_{LLL} and X_{LLH}, and detail coefficients C_{detail} from the Embedding module, reconstructing details via 3D-IDWT and learnable 3D Conv.

Wavelet Bottleneck. This module is located in the middle stage achieving via translators and wavelet transform. Each wavelet bottleneck block first decomposes feature maps X_{in} into approximation C_{a} and detail C_{d} coefficients. C_{a} is then passed as input to translators for deeper processing. The processed C_{a} and the original detail subbands C_{d} are transformed back to the original space via 3D-IDWT, which allows low-frequency features to be focused while retaining high-frequency details.

Wavelet-Domain High-Frequency Focal Loss

The traditional mean squared error (MSE) loss treats each pixel equally, which makes it more emphasizes low-frequency supervision. To compensate for the inadequacies of high-frequency supervision in MSE, we propose the wavelet-domain High-Frequency Focal Loss (HFFL) to supervise high-frequency variations.

Focal Frequency Loss (FFL) (Jiang et al. 2021) aimed to decrease the frequency distance between real and generated

images. It can be defined as:

$$\mathcal{L}_{\text{FFL}} = \frac{1}{HW} \sum_{u=0}^{H-1} \sum_{v=0}^{W-1} w(u,v) \left| F_r(u,v) - F_f(u,v) \right|^2, \quad \text{where } w(u,v) = \left| F_r(u,v) - F_f(u,v) \right|^{\alpha}, \tag{9}$$

The weight matrix $w(u,v)$ is guided by a frequency-specific training loss. However, this approach inadequately reflects high-frequency discrepancies due to the large dynamic frequency range. Thus we refine the weight matrix of FFL by incorporating the absolute value of the logarithm $|\log(w(u,v))|$ for the high-frequency band while low-frequency weights are set to zero via a predefined threshold τ. We further refine the frequency-domain loss in a multi-scale manner, we adopt the wavelet transform to decompose the signal into multi-level subbands, the wavelet-domain high-frequency focal loss can be defined as:

$$\mathcal{L}_{\text{HFFL}} = \frac{1}{H_k W_k} \sum_{u=0}^{H_k-1} \sum_{v=0}^{W_k-1} w(u,v) \left| F_r(u,v) - F_f(u,v) \right|^2, \quad w(u,v) = \left| \log\left(\left| F_r(u,v) - F_f(u,v) \right| \right) \right|^{\alpha}, \tag{10}$$

where k denotes the k-level wavelet decomposition. Our model is trained end-to-end with the objective function combining spatial- and wavelet-domain losses for each frame:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \lambda \mathcal{L}_{\text{HFFL}}. \tag{11}$$

where MSE loss is the spatial domain loss, and λ is a weight parameter to balance the two losses.

Experiments

In this section, we present detailed experimental results. We evaluate our proposed method against strong recent baselines, including competitive recurrent-based architectures: ConvLSTM (Shi et al. 2015), PredRNN (Wang et al. 2017), PredNet (Lotter et al. 2016), PredRNN++ (Wang et al. 2018a), MIM (Wang et al. 2019), E3D-LSTM (Wang et al. 2018b), PhyDNet (Guen et al. 2020), MAU (Chang et al. 2021), and PredRNNv2 (Wang et al. 2022). We also compare recent recurrent-free architectures: SimVP (Gao et al. 2022) and TAU (Tan et al. 2023a).

Experiment Setting

Multi-Scenario. Our model is quantitatively evaluated across varied real-world scenarios, including driving scenes, human motion capture, traffic flow prediction, and weather forecasting, encompassing micro to macro scales. Dataset statistics are summarized in Table 1.

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
WeatherBench	2,167	706	1	32	64	12	12

Table 1: The detailed summary of the dataset statistic. The number of samples, input frames T, and predicted frames T' are shown for the training and testing sets.

Evaluation Metrics. The model performance in various scenarios is evaluated through multiple metrics. *Error metrics*, including mean squared error (MSE), mean absolute error (MAE), and root mean squared error (RMSE), measure pixel-wise error. *Similarity metrics* include the structural similarity index measure (SSIM) and peak signal-to-noise ratio (PSNR). *Computational metrics* such as the number of parameters and floating-point operations (FLOPs) also evaluate the models.

Implementation Details. Our proposed method is implemented in Pytorch and conducts experiments on a single NVIDIA-V100 GPU. The model trained with a mini-batch of 16 video sequences, employs the AdamW optimizer, OneCycle learning rate scheduler, and weight decay of $5e^{-2}$. Optimal learning rate is chosen from $\left\{1e^{-2}, 5e^{-3}, 1e^{-3}\right\}$ for stable training. We utilize stochastic depth for regularization to avoid overfitting.

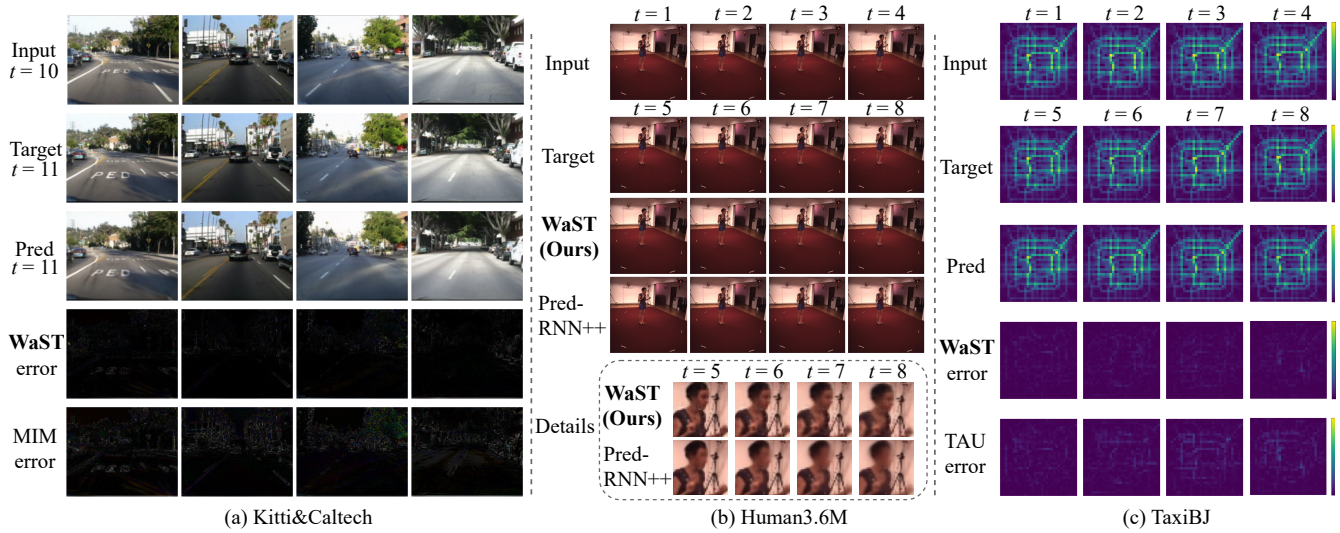
Driving Scenes Prediction

Kitti&Caltech Generalization capability is crucial in artificial intelligence, often challenged in traditional supervised learning across diverse domains. Self-supervised learning strives to construct robust representation from unlabeled data, with the generalization ability evaluated based on the learned model through downstream tasks. We evaluate this capability across different datasets, training the model on KITTI (Geiger et al. 2013) and testing on Caltech Pedestrian (Dollár et al. 2009).

Quantitative results Table 2 demonstrate our model achieves state-of-the-art performance under all metrics. Notably, our approach surpasses the prior state-of-the-art MIM method (Wang et al. 2019), attaining a 63-fold computational reduction (1858.0G $\rightarrow$ 29.4G) and a 5.7-fold parameter reduction (49.2M $\rightarrow$ 8.6M). Qualitative visualizations Figure 5(a) present that our method learns spatiotemporal dependencies from past frames and predicts high-quality future frames. From the final rows in Figure 5(a), it becomes evident that the recurrent-based MIM (Wang et al. 2019) struggles to pinpoint distant, small entities. The robustness of our approach in variations of illumination and lane lines indicates the potential value for autonomous vehicles.

Kitti&Caltech (10 $\rightarrow$ 1 frames)						
Method	Date	FLOPs (G)	MSE $\downarrow$	MAE $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
ConvLSTM	NIPS'2015	595.0	139.6	1583.3	0.9345	27.46
PredRNN	NIPS'2017	1216.0	130.4	1525.5	0.9374	27.81
PredRNN++	ICML'2018	1803.0	125.5	1453.2	0.9433	28.02
MIM	CVPR'2019	1858.0	125.1	1464.0	0.9409	28.10
E3D-LSTM	ICLR'2019	1004.0	200.6	1946.2	0.9047	25.45
PhyDNet	CVPR'2020	40.4	312.2	2754.8	0.8615	23.26
MAU	NIPS'2021	172.0	177.8	1800.4	0.9176	26.14
SimVP	CVPR'2022	60.6	160.2	1690.8	0.9338	26.81
PredRNNv2	PAMI'2022	1223.0	147.8	1610.5	0.9330	27.12
TAU	CVPR'2023	92.5	131.1	1507.8	0.9456	27.83
Ours	-	29.4	**123.5**	**1384.2**	**0.9468**	**28.49**

Table 2: Quantitative results in Kitti&Caltech dataset.



(a) Kitti&Caltech (b) Human3.6M (c) TaxiBJ

Figure 5: Qualitative visualizations across varied real-world datasets, including (a) Kitti&Caltech, (b) Human3.6M, and (c) TaxiBJ datasets, where prediction error = |target - prediction|.

Human Motion Capture

Human3.6M Human motion prediction brings inherent challenges due to the high resolution and complex dynamics associated with unpredictable human behavior. Traditional recurrent-based methods struggle to predict detailed features of moving human bodies, particularly the face and hands, due to their limited pixel representation.

Our proposed model, despite dealing with higher resolution, exhibits state-of-the-art performance across all metrics, as outlined in Table 3. Remarkably, our approach outperforms the prior state-of-the-art PredRNN++ (Wang et al. 2018a), securing a 21-fold computational reduction (1033.0G $\rightarrow$ 49.3G) and a 3.7-fold parameter reduction (39.3M $\rightarrow$ 10.6M). The adaptability of our method is superior to the recurrent unit in capturing dynamic scenes due to the unique design of the frequency and time attention mechanisms. The visualization is presented in Figure 5(b), the prediction details demonstrate that our method dynamic changes over time without losing human detail information, implying potential applicability in complex dynamic scenes.

Method	Date	Human3.6M ($4 \rightarrow 4$ frames) FLOPs (G)	MSE $\downarrow$	MAE $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
ConvLSTM	NIPS'2015	347.0	125.5	1566.7	0.9813	33.40
PredNet	ICLR'2017	13.7	261.9	1625.3	0.9786	31.76
PredRNN	NIPS'2017	704.0	113.2	1458.3	0.9831	33.94
PredRNN++	ICML'2018	1033.0	110.0	1452.2	0.9832	34.02
MIM	CVPR'2019	1051.0	112.1	1467.1	0.9829	33.97
E3D-LSTM	ICLR'2019	542.0	143.3	1442.5	0.9803	32.52
MAU	NIPS'2021	105.0	127.3	1577.0	0.9812	33.33
SimVP	CVPR'2022	197.0	115.8	1511.5	0.9822	33.73
PredRNNv2	PAMI'2022	708.0	114.9	1484.7	0.9827	33.84
TAU	CVPR'2023	182.0	113.3	1390.7	0.9837	34.03
Ours	-	49.3	**109.8**	**1384.7**	**0.9839**	**34.19**

Table 3: Quantitative results in Human3.6M dataset.

Traffic Flow Prediction

TaxiBJ Spatiotemporal traffic flow prediction is challenging due to complexities stemming from human behaviors. Evaluating our model on the real-world traffic dataset TaxiBJ (Zhang et al. 2017). Traditional forecasting methods struggle with intricate road network interdependencies and non-linear temporal dynamics.

The quantitative results are reported in Table 4 and qualitative visualizations in Figure 5(c). Our model consistently generates precise predictions, despite subtle frame differences. This stems from an oversized convolutional kernel (51×51), larger than the image resolution (32×32), adept at low-frequency feature capture, resulting in near-perfect prediction errors (Figure 5(c), last two rows). Our model notably surpasses previous methodologies across all metrics (Table 4), maintaining modest computational complexity (1.0G), and indicating practical applicability.

Method	Date	TaxiBJ ($4 \rightarrow 4$ frames) FLOPs (G)	MSE $\times 100\downarrow$	MAE $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$
ConvLSTM	NIPS'2015	20.7	48.5	17.7	0.978	37.38
PredRNN	NIPS'2017	42.4	46.4	17.1	0.971	38.52
PredRNN++	ICML'2018	63.0	44.8	16.9	0.977	38.71
MIM	CVPR'2019	64.1	42.9	16.6	0.971	38.71
E3D-LSTM	ICLR'2019	98.2	43.2	16.9	0.979	38.75
SimVP	CVPR'2022	3.6	41.4	16.2	0.982	39.17
PredRNNv2	PAMI'2022	42.6	38.3	15.6	0.983	39.38
TAU	CVPR'2023	2.5	34.4	15.6	0.983	39.50
Ours	-	1.0	**30.8**	**14.9**	**0.984**	**39.73**

Table 4: Quantitative results in the TaxiBJ dataset.

Weather Forecasting

WeatherBench Climate prediction, a crucial task in spatiotemporal predictive learning, emphasizes the need for ro-

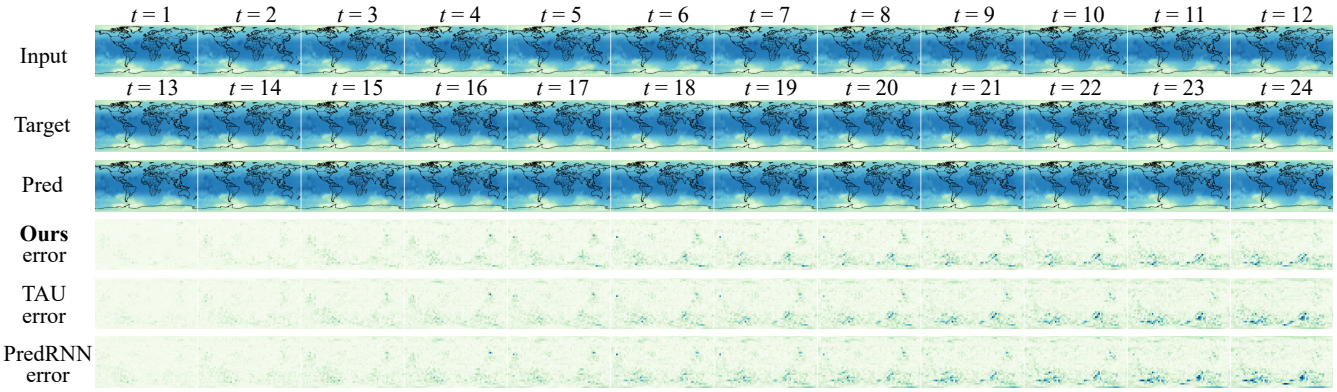


Figure 6: Qualitative visualizations on global temperature forecasting in WeatherBench. The differences between the ground truth and the predicted frames are visualized in the last three rows.

bust data-driven models over purely physical methods, due to high computational demands. In this context, we compare MSE, MAE, and RMSE metrics for temperature prediction at $5.625°$ resolution (32×64 grid points) as reported in Table 5. Our time-frequency aware design successfully captures low and high frequencies in climatic factors, providing qualitative global temperature visualization Figure 6. It surpasses existing climate prediction models, notably enhancing the state-of-the-art recurrent-based PredRNN (Wang et al. 2017) performance in MSE (1.331 $\rightarrow$ 1.098), and reducing computational costs and parameters by 185-fold (278.0G $\rightarrow$ 1.5G) and 6-fold (23.6M $\rightarrow$ 3.9M). This demonstrates the effectiveness of our method for global climate forecasting.

Method	WeatherBench (12 $\rightarrow$ 12 frames) Date	FLOPs (G)	MSE ↓	MAE ↓	RMSE ↓
ConvLSTM	NIPS'2015	136.0	1.521	0.7949	1.233
PredRNN	NIPS'2017	278.0	1.331	0.7246	1.154
PredRNN++	ICML'2018	413.0	1.634	0.7883	1.278
MIM	CVPR'2019	109.0	1.784	0.8716	1.336
E3D-LSTM	ICLR'2019	169.0	1.592	0.8059	1.233
MAU	NIPS'2021	39.6	1.349	0.7391	1.162
SimVP	CVPR'2022	8.0	1.238	0.7037	1.113
PredRNNv2	PAMI'2022	279.0	1.545	0.7986	1.243
TAU	CVPR'2023	6.7	1.224	0.6810	1.106
Ours	-	1.5	**1.098**	**0.6338**	**1.044**

Table 5: Quantitative results on the temperature forecasting in WeatherBench dataset.

Ablation Study

In this section, we ablate essential design choices in WaST on WeatherBench humidity and cloud cover prediction. We compare the recurrent-based counterpart with our proposed wavelet-based recurrent-free architecture. We also compare the proposed Time-Frequency Aware Translator and advanced MetaFormer (Yu et al. 2022) modules. The baseline is obtained by copying the input as the prediction.

In Table 6 details, two key observations emerge: (i) Recurrent-free architectures surpass recurrent-based ones, attributed to the enhanced spatiotemporal learning capacities of translators. (ii) Compared with various MetaFormer modules, our proposed TF-Aware Translator stands out in performance. Additionally, models trained without HFFL (w/o HFFL) exhibit diminished prediction accuracy, underscoring the importance of high-frequency information supervision.

Method		Humidity MAE↓	Humidity RMSE↓	Cloud Cover MAE↓	Cloud Cover RMSE↓
Baseline	Copying	9.046	13.346	0.2156	0.3361
Recurrent based	MIM	5.504	7.817	0.1718	0.2449
	PhyDNet	8.975	15.460	0.2261	0.3149
	E3DLSTM	4.100	6.044	0.1529	0.2394
	PredRNN	4.096	6.133	0.1588	0.2346
	PredRNN++	4.731	6.782	0.1544	0.2341
Recurrent free	ViT	3.911	5.742	0.1503	0.2186
	Uniformer	3.914	5.734	0.1485	0.2170
	MLP-Mixer	3.950	5.871	0.1526	0.2219
	ConvMixer	3.909	5.730	0.1487	0.2172
	ConvNeXt	3.928	5.760	0.1487	0.2178
	HorNet	3.906	5.721	0.1481	0.2174
	Ours w/o HFFL	3.721	5.600	0.1469	0.2167
	Ours	**3.694**	**5.569**	**0.1452**	**0.2150**

Table 6: Ablation study on WeatherBench dataset.

Conclusion

This paper presents an innovative wavelet-based spatiotemporal prediction framework WaST, leveraging the wavelet transform across image and feature levels. Our Time-Frequency Aware Translator adaptively modulates both low and high-frequency components in multi-level wavelet space. The wavelet-domain High-Frequency Focal Loss is introduced to compensate for the inadequacies of traditional mean squared error in high-frequency supervision. Overall, our approach emphasizes the significance of both intra-frame frequency and inter-frame temporal variations, enabling the model to capture short- and long-term information. Demonstrably, our method provides state-of-the-art performance on various datasets with real-world scenarios.

References

Bae, W.; ; et al. 2017. Beyond deep residual learning for image restoration: Persistent homology-guided manifold simplification. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 145–153.

Chang, Z.; Zhang, X.; Wang, S.; Ma, S.; Ye, Y.; Xinguang, X.; and Gao, W. 2021. Mau: A motion-aware unit for video prediction and beyond. *Advances in Neural Information Processing Systems*, 34: 26950–26962.

Ding, X.; Zhang, X.; Han, J.; and Ding, G. 2022. Scaling up your kernels to 31x31: Revisiting large kernel design in cnns. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11963–11975.

Dollár, P.; Wojek, C.; Schiele, B.; and Perona, P. 2009. Pedestrian detection: A benchmark. In *2009 IEEE conference on computer vision and pattern recognition*, 304–311. IEEE.

Fang, S.; Zhang, Q.; Meng, G.; Xiang, S.; and Pan, C. 2019. GSTNet: Global Spatial-Temporal Network for Traffic Flow Prediction. In *IJCAI*, 2286–2293.

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

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Guen, V. L.; et al. 2020. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11474–11484.

Guo, T.; Seyed Mousavi, H.; Huu Vu, T.; and Monga, V. 2017. Deep wavelet prediction for image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 104–113.

Gupta, A.; Tian, S.; Zhang, Y.; Wu, J.; Martín-Martín, R.; and Fei-Fei, L. 2022. Maskvit: Masked visual pre-training for video prediction. *CVPR*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hochreiter, S.; et al. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Hu, J.; et al. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Hu, X.; Huang, Z.; Huang, A.; Xu, J.; and Zhou, S. 2023. A dynamic multi-scale voxel flow network for video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6121–6131.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021. Focal frequency loss for image reconstruction and synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13919–13929.

Li, Q.; Shen, L.; Guo, S.; and Lai, Z. 2020. Wavelet integrated CNNs for noise-robust image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7245–7254.

Li, S.; Wang, Z.; Liu, Z.; Tan, C.; Lin, H.; Wu, D.; Chen, Z.; Zheng, J.; and Li, S. Z. 2022. Efficient Multi-order Gated Aggregation Network. *ArXiv*, abs/2211.03295.

Li, S.; Wu, D.; Wu, F.; Zang, Z.; and Li, S. Z. 2023. Architecture-Agnostic Masked Image Modeling – From ViT back to CNN. In *International Conference on Machine Learning*.

Liu, P.; Zhang, H.; Zhang, K.; Lin, L.; and Zuo, W. 2018. Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 773–782.

Liu, S.; Chen, T.; Chen, X.; Chen, X.; Xiao, Q.; Wu, B.; Pechenizkiy, M.; Mocanu, D.; and Wang, Z. 2023. More convnets in the 2020s: Scaling up kernels beyond 51x51 using sparsity. *ICLR*.

Lotter, W.; et al. 2016. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*.

Nie, X.; Chen, X.; Jin, H.; Zhu, Z.; Yan, Y.; and Qi, D. 2023. Triplet Attention Transformer for Spatiotemporal Predictive Learning. *arXiv preprint arXiv:2310.18698*.

Oyallon, E.; et al. 2017. Scaling the scattering transform: Deep hybrid networks. In *Proceedings of the IEEE international conference on computer vision*, 5618–5627.

Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; and Prabhat, f. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743): 195–204.

Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.

Tan, C.; Gao, Z.; Li, S.; and Li, S. Z. 2022. Simvp: Towards simple yet powerful spatiotemporal predictive learning. *arXiv preprint arXiv:2211.12509*.

Tan, C.; Gao, Z.; Wu, L.; Xu, Y.; Xia, J.; Li, S.; and Li, S. Z. 2023a. Temporal attention unit: Towards efficient spatiotemporal predictive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18770–18782.

Tan, C.; Li, S.; Gao, Z.; Guan, W.; Wang, Z.; Liu, Z.; Wu, L.; and Li, S. Z. 2023b. OpenSTL: A Comprehensive Benchmark of Spatio-Temporal Predictive Learning. *arXiv preprint arXiv:2306.11249*.

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

Wang, Y.; Jiang, L.; Yang, M.-H.; Li, L.-J.; Long, M.; and Fei-Fei, L. 2018b. Eidetic 3D LSTM: A model for video prediction and beyond. In *International conference on learning representations*.

Wang, Y.; Long, M.; Wang, J.; Gao, Z.; and Yu, P. S. 2017. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Advances in neural information processing systems*, 30.

Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Philip, S. Y.; and Long, M. 2022. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2208–2225.

Wang, Y.; Zhang, J.; Zhu, H.; Long, M.; Wang, J.; and Yu, P. S. 2019. 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*, 9154–9162.

Xu, Z.; Liu, Z.; Sun, C.; Murphy, K.; Freeman, W. T.; Tenenbaum, J. B.; and Wu, J. 2019. Unsupervised discovery of parts, structure, and dynamics. *ICLR*.

Yao, T.; Pan, Y.; Li, Y.; Ngo, C.-W.; and Mei, T. 2022. Wave-vit: Unifying wavelet and transformers for visual representation learning. In *ECCV*, 328–345. Springer.

Yu, W.; Luo, M.; Zhou, P.; Si, C.; Zhou, Y.; Wang, X.; Feng, J.; and Yan, S. 2022. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10819–10829.

Zhang, J.; et al. 2017. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.