Exploring Spatial Frequency Information for Enhanced Video Prediction Quality

Junyu Lai, *Member, IEEE*, Lianqiang Gan, Junhong Zhu, Huashuo Liu, and Lianli Gao, *Member, IEEE*

***Abstract*—Video prediction is a challenging spatiotemporal prediction task that generates future frames based on historical observations. Although recently proposed deep learning-based methods significantly outperform legacy approaches, there still exist gaps between prediction and ground truth, primarily rooted in edge and motion blurring. On the one hand, since conventional performance metrics like Mean Square Error (MSE) and Structure Similarity Index Measure (SSIM) cannot decently evaluate this deficiency, we design a 3D Frequency Loss (3DFL) metric to better assess the similarity of predicted video frames. On the other hand, edge and motion blurring is mainly attributed to the predictive model's insufficient attention to high spatial frequency arising from rapid pixel value variations at object edges, and it is observed that shallow networks are more adept at capturing high spatial frequency information. Therefore, aiming to alleviate edge and motion blurring, we proposed a novel video prediction model termed SDFNet that can extract and integrate both spatially encoded shallow and deep-level features. To accommodate SDFNet's multi-branch input structure, a frequency adaptive translator (FATranslator) is derived, which leverages involution operators to adaptively extract inter-frame temporal dependencies from different spatial encoding layers, and further mitigates motion blurring. Extensive experiments demonstrate that our proposed model achieves significant improvements in prediction accuracy and temporal consistency over the current state-of-the-art models on various benchmarks. The results highlight the importance of spatial frequency modeling for enhancing video prediction performance.**

***Index Terms*—Video prediction, deep learning, spatial frequency, shallow and deep-level features, performance metric.**

I. INTRODUCTION

VIDEO prediction, a technique that generates future frames based on past input sequences, has garnered significant research interest in various domains such as traffic flow prediction [1], weather nowcasting [2], [3] , human action forecasting [4], [5], [6], and autonomous vehicles [7]. Given its potential to predict complex scenes with more than three dimensions of data, video prediction presents a highly challenging task in mathematical modeling. Deep learning-based methods have shown promise in tackling this challenge, as they can effectively learn representations from high-dimensional data [8], enabling unsupervised learning for predicting intricate nature videos.

Despite deep learning-based methods significantly outperform legacy approaches, edge and motion blurring occurred in video frames predicted by state-of-the-art (SOTA) models seriously influenced video prediction quality. On the one hand, edge blur pertains to the absence of intricate details within regions of pronounced contrast, such as object boundaries or varying textures, in video frames. On the other hand, motion blur refers to the blurring and loss of detail of moving objects in the predicted frames, caused by inaccurate trajectory and motion estimation. Both edge and motion blurs can result in the challenge of accurately reproducing sharp edges and intricate texture details of moving objects across frames. Many recent research efforts have already discussed the deficiency. Xu *et al.* [9] discovered prior model tends to produce blurry results, which compromised the quality of video prediction. Oprea *et al.* [10] explicated pointed out the widely used performance metric such as MSE and SSIM cannot decently measure edge and motion blurring. Jin *et al.* [11] proposed a model that embed digital image processing algorithms to address the problem of blurry prediction. However, few practical solutions tackling with this deficiency are presented in the aforementioned literature. Therefore, in this paper, we aim to solve this substantial problem to further enhance video prediction quality.

Commonly employed predictive deep neural network (DNN) models for video prediction encompass architectures such as autoencoders [12], convolutional neural networks (CNNs) [13], recurrent neural networks (RNNs) [5], [14], and generative adversarial networks (GANs) [15]. To enhance prediction accuracy, these networks are often combined in various ways, such as the formations of CNN-RNN-CNN, CNN-CNN-CNN, among others [16]. These integrated models typically consist of three main components: spatial feature extraction, temporal feature extraction, and feature fusion. Each component is designed to adaptively capture the spatial and temporal features present in the data. Consequently, these models have demonstrated superior performance in numerous video prediction tasks. However, challenges persist when predicting videos containing chaotic scenes and complex dynamics, making it difficult to accurately extract

The next few paragraphs should contain the authors' current affiliations, including current address and e-mail. For example, First A. Author is with the National Institute of Standards and Technology, Boulder, CO 80305 USA (e-mail: author@ boulder.nist.gov). *(Corresponding author: Lianli Gao.)*

Junyu Lai is with the School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, and also with the Aircraft Swarm Intelligent Sensing and Cooperative Control Key Laboratory of Sichuan Province, Chengdu 611731, China (e-mail: laijy@uestc.edu.cn).

Lianqiang Gan, Junhong Zhu and Huashuo Liu are with the School of Aeronautics and Astronautics, University of Electronic Science and Technology of China, Chengdu 611731, China, (e-mail: laijy@uestc.edu.cn; ganlq@std.uestc.edu.cn; zhujh@std.uestc.edu.cn; liuhs@std.uestc.edu.cn).

Lianli Gao is with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, (e-mail: gaoll@uestc.edu.cn).

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

spatiotemporal information and finally generate well-predicted video frames.

To improve video prediction accuracy in chaotic scenarios, improving the model's perception of inconsistent dynamics becomes imperative. Optical flow warping [17], [18] and inter-frame differential [19] are two main approaches to enhance motion information in video frames. However, computing optical flow for input frames is computationally intensive and time-consuming, leading to a longer training and inference time. Differential operation is another method to capture motion information by comparing pixel value changes between adjacent frames, which also incurs heavy computation costs. Additionally, incorporating channel attention mechanisms [20], [21], [22], [23] can mitigate background interference and emphasize dynamic information. While successful in image domain tasks, models with attention mechanisms show limitations in modeling temporal tasks including video prediction. Moreover, attention-based models often require substantial data and training time, restricting their applications. Strengthening spatiotemporal information fusion is another solution. For instance, 3D convolution operation can jointly extract both temporal and spatial information, making it more suitable for spatiotemporal sequence tasks [24], [25]. Nevertheless, 3D convolution's high parameter overhead hampers its scalability. Despite the recent progress made by the existing methods, gaps still exist between prediction and ground truth. Specifically, a prominent inconsistency lies in edge and motion blurring.

We also note that an appropriate performance metric for video prediction is of utmost importance. Commonly used metrics in video prediction are mainly adapted from the image domain, such as pixel-wise metrics like mean square error (MSE) and mean absolute error (MAE). Although effective in measuring prediction accuracy for some synthetic datasets with deterministic scenarios, MSE and MAE struggle with more practical videos containing intricate and uncertain backgrounds. Hence, structural metrics are predominantly utilized to evaluate video prediction accuracy, such as the structure similarity index measure (SSIM) [26] and peak signal-to-noise ratio (PSNR). Compared to pixel-wise metrics, structural metrics consider the pixel distribution structure of images, effectively measuring image similarity. However, in the context of video prediction, SSIM and PSNR may not always align well with human visual perception. This inconsistency will be demonstrated in Section III, where human perceived edge and motion blurring cannot be objectively measured by SSIM and PSNR.

In this paper, we firstly introduce a novel metric called 3D frequency loss (3DFL) based on 3D fast Fourier transform (3DFFT). Extending traditional 2D methods, 3DFL jointly captures spatial pixel distributions within frames and temporal pixel variations between consecutive frames, enabling more comprehensive assessment for edge and motion blurring in predictive video frames. On the other hand, it is believed that the primary cause of edge and motion blurring lies in the predictive model's limited focus on high spatial frequency information resulting from rapid pixel value variations at object edges, and we observe shallow networks have shown greater proficiency in capturing high spatial frequency information. Therefore, to address the challenge of edge and motion blurring, we elaborate an end-to-end video prediction model, termed SDFNet, which can extract and seamlessly incorporate spatially encoded shallow and deep-level features. To effectively accommodate SDFNet's multi-branch input structure, we formulate a frequency adaptive translator (FATranslator) comprising specialized blocks tailored to each input scale. By incorporating involution operators with dynamic kernels and lower overhead in lieu of traditional convolutions, FATranslator can selectively emphasize different frequency components, and improves SDFNet's ability to learn inter-frame temporal correlations. Our proposal aims to significantly enhance prediction quality by effectively exploring spatial frequency information, temporal correlation, and finally mitigating the blurring phenomenon.

To summarize, the main contributions of this paper are:

1) We introduce 3DFL, a new performance metric that effectively combines pixel distribution intra-frame and pixel variation inter-frames using 3DFFT, making it particularly suitable for measuring edge and motion blurring in video prediction.
2) We propose an innovative SDFNet video prediction model, which can extract and integrate deep spatial features with the often-overlooked shallow spatial features in legacy models, resulting in the generation of more accurate and less blurring predictions.
3) We elaborate a multi-layered FATranslator incorporating involution operators to enhance the inter-frame temporal correlation learning capability, thus to further mitigate motion blurring in predicted frames.
4) We conduct comprehensive experiments to evaluate SDFNet's performance on diverse benchmarks. The results unequivocally showcase remarkable improvements in prediction fidelity and temporal consistency when compared to baseline approaches.

The remaining part of this paper proceeds as follows: Section II reviews related works. Section III empirically analyzes existing problems and reveals the limitations of legacy metrics and models. Section IV presents and validates the proposed 3DFL metric. Section V elaborates the SDFNet model, which can extract and integrate spatially encoded shallow and deep-level features. Section VI provides extensive experiments on multiple datasets demonstrating SDFNet's superior performance over baselines. Finally, Section VII concludes with discussions and directions for future research.

II. RELATED WORKS

In this section, we review the existing research efforts related to video frame prediction, covering four aspects: Video predictive learning, video prediction metric, involution operation, and frequency analysis in video prediction.

A. Video Predictive learning

In recent years, researchers have particularly focused on

improving RNN models for video prediction tasks, given their excellent performance in time series prediction. Several noteworthy advancements have emerged from this line of research. ConvLSTM [2] extended legacy LSTM by replacing the fully connected layer with a convolutional layer to process spatial information, demonstrating a typical case for subsequent spatiotemporal prediction works. Inspired by ConvLSTM, PredRNN [27] proposed the ST-LSTM unit, which employs three memory gates to control memory flow, enabling simultaneous memorization of spatial and temporal states. PredRNN++ [28] further introduced the gradient highway unit and Casual LSTM to adaptively capture temporal dependencies. PredRNNv2 [29] improved PredRNN by introducing a reverse scheduled sampling method. E3D-LSTM [25] merged 3D convolution with RNN to enhance the ability to extract multidimensional features. PhyDnet [30] employed partial differential equations to design the Phycell, enabling the learning of known physical dynamics in the latent space. CrevNet [31] leveraged reversible architecture to build a bijective two-way autoencoder and its complementary recurrent predictor for spatiotemporal predictive learning. SimVP [16] is an innovative pure CNN-based model that effectively extracts temporal and spatial features, demonstrating the suitability of CNN models for temporal extraction. Moreover, many works have integrated the attention mechanism to improve model performance in video prediction tasks. For example, LMC-memory [32] proposed a long-term motion context memory that combines the attention mechanism with ConvLSTMs. MAFE [22] designed a Channel-wise and Spatial Attention (CSA) module to enhance motion-aware features.

B. Video Prediction Metric

Current video predictive models [2], [16], [25], [27]-[32] commonly employ image similarity metrics, including Mean MSE, MAE, SSIM, PSNR, and Learned Perceptual Image Patch Similarity (LPIPS) to evaluate prediction accuracy. However, as images are two-dimensional data with strong correlations between pixels, MSE and MAE, which calculate absolute pixel differences, may not sufficiently assess image similarity. To address this limitation, Zhou et al. [33] introduced SSIM, which considers brightness, contrast, and structural information to calculate a similarity index value between 0 and 1. Besides, PSNR is a commonly used metric for measuring the similarity between an original image and a processed image, with a higher value indicating higher similarity and lower distortion between the images. LPIPS [34] also measures image similarity using DNN models that mimic the human visual system, accounting for perceptual effects like color, texture, and structure to better match human perception. However, these metrics evaluate videos as simple collections of frames, overlooking inter-frame correlations. To better evaluate the accuracy of video prediction, particular in alignment with human visional perception, novel performance metrics need to be derived.

C. Involution Operation

To improve the ability of temporal dependency learning without solely relying on RNN, CNN has also been utilized for video prediction [12], [13], [16]. The involution operation, previously employed in RedNet [35] for tasks such as classification, detection, and segmentation, has demonstrated its effectiveness in surpassing convolutional operations. Compared to traditional convolution, involution reduces the number of parameters and computational costs. More specifically, involution generates kernels using self-attention mechanisms and performs multiplication and addition with the input. This adaptability in kernel generation helps match the input data better, and as a result facilitates more effective extraction of inter-frame temporal correlations. Furthermore, involution maintains the same receptive field as traditional convolution while significantly reducing the number of model parameters, making it a promising addition to video prediction tasks. Therefore, we introduce involution operator into video prediction for enhancing model accuracy and efficiency, simultaneously.

D. Frequency Analysis in Video Prediction

Frequency analysis in digital image processing has proven common and efficient. In video prediction, frequency analysis helps analyze errors and guide model design. Some previous works have incorporated frequency analysis into video prediction. Jin *et al*. [11] emphasized the importance of frequency analysis for spatial and temporal frequency signals, leading to the design of DWT-S and DWT-T on temporal and spatial axes, respectively. Xu *et al*. [9] preprocessed video frames with a high-pass filter to obtain high frequency information, which are then fed into two different encoders along with the original frames. Farazi *et al*. [36] calculated the phase difference between Fourier transformed results for continuous frames, interpreted as a local offset, and spatiotemporally adjusted by the transform model.

However, embedding digital image processing algorithms into DNN models may introduce computational burden and significant time consumption. In order to integrate frequency analysis with DNN without resorting to digital signal processing algorithms, it is essential to investigate the characteristics of DNN models in fitting spatial frequency information. There are many previous efforts devoted to studying the relationship between spatial frequency and DNNs. Wang *et al*. [37] discovered that high spatial frequency components contribute to the generalization of DNNs. F-principle [38] and Deep frequency principle [39], highlighted that deeper neural network tends to fit low spatial frequency information better. Huang *et al*. [40] designed a deeper neural network to remove high spatial frequency noise in video. Consequently, shallow features along with high spatial frequency information deserve more attention. Combining shallow features can enable the model to pay more attention to high spatial frequency information, alleviating edge and motion blurring in video prediction.

III. Existing Problems

To demonstrate the visualization inconsistency of conventional performance metrics and the blurring deficiency of existing video prediction models, we present a comprehensive analysis based on visualization experiments using representative models including SimVP, PredRNNv2, PredRNN, and ConvLSTM. The experiments are conducted on the KTH [41] dataset, a human actions dataset depicting various poses like waving and jogging in different scenarios (cf. Section VI). This dataset contains varying scenarios and complex human actions, which leads traditional metrics often yield evaluations that do not align with human perception. Specifically, these selected models are trained to predict the next 20 frames based on the previous 10 frames. As exemplified in Fig. 1, each row represents various examples, and each column represents distinct models, except for the first column, which is the ground truth. While achieving reasonable MSE and SSIM scores, the predicted frames exhibit considerable blurring compared to the ground truth.

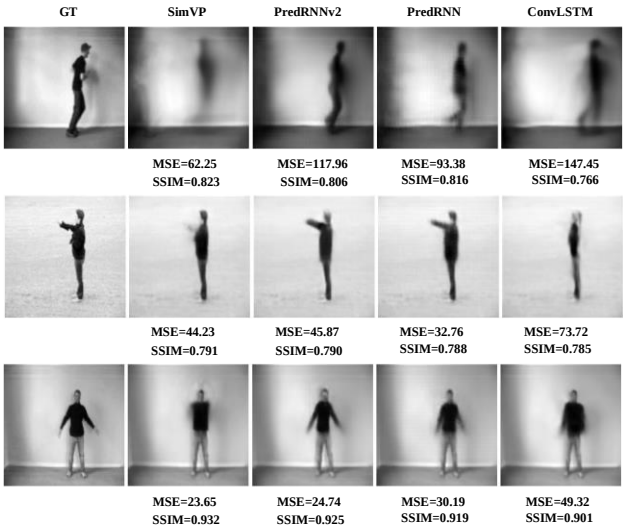

Fig. 1. Visualization experiment of four chosen baselines. MSE and SSIM metrics are labeled below figures.

- In the first example, the person is jogging, exhibiting various motion features. SimVP's predictions suffer from serious blurriness, but it achieves the best performance on MSE and SSIM. PredRNN and PredRNNv2, despite having lower MSE and SSIM values, generate clear frames consistent with human perception.
- In the second and the third examples, the person is boxing and waving, respectively. Their arms contain more motion information while the body remains static. SimVP accurately predicts the static body but struggles with the moving arms. PredRNN and PredRNNv2 perform well in modeling both the moving arms and static body with higher clarity and consistency. However, SimVP still outperforms PredRNN in MSE and SSIM metrics, which contradicts the visualization results.

Based on the above analysis, it becomes evident that traditional MSE and SSIM metrics are inconsistent with the visualization of prediction results and cannot precisely measure the quality of video prediction. Therefore, it is imperative to propose a new metric to measure human perceived edge and motion blurring in predicted video frames.

Furthermore, by conducting a cross-model comparison, we find that PredRNN, PredRNNv2, and ConvLSTM, whose temporal feature extractors comprise RNNs outperform CNN-based SimVP in the visualization experiment. For example, PredRNN's predicted results aligning better with human visual perception. This is because when a RNN learns temporal features, each frame receives hidden states from the previous frames, resulting in close relationships between the predicted results of the frames. In contrast, CNN tends to take a localized approach, focusing more on extracting short-term patterns from video sequences. This leads to blurry predicted frames as CNN has difficulty modeling long-term temporal dependencies. Therefore, research on video prediction models, which can generate more accurate and less blurring predictions is vital.

IV. Metric Design and Validation

A. *Metric design*

Traditional metrics like MSE and MAE, which are solely based on pixel value differences, tend to blur predictions due to their accommodation of uncertainty [10], When the background dominates a frame, they are less affected by small moving objects. Similarly, SSIM also loses focus on moving objects when a high proportion of background is present. Furthermore, since a video consists of multiple frames with strong temporal correlations, traditional metrics evaluate video quality by treating each frame independently, neglecting inter-frame temporal correlation. Therefore, a new metric is required to capture both the spatial frequency information within a frame and the temporal correlation across frames.

Capture Intra-frame Spatial Frequency Information: A feasible approach is using the 2D Fourier transform (2DFFT) to process frames, which separates edge parts with rapidly changing pixel values from background parts with slowly variable pixel values. This processing results in the separation of edges and background into high-frequency and low-frequency information, respectively. The blurriness at the edges becomes clearly reflected as high-frequency information loss. This approach transforms video frames into the frequency domain, allowing the edge blurriness problem undetected by MSE and SSIM to be more intuitively reflected.

Capture Inter-frame Temporal Correlation: Using 2DFFT alone still overlooks temporal dynamics and motion correlations between frames. To address object motion blur, it is crucial to enable temporal consistency between frames. This can be solved by adding the temporal dimension to 2DFFT. Notably, motion information in the temporal dimension can also be described using Fourier transform. For instance, in a hand-waving action, fast-moving parts like the arms are classified as high-frequency information, while slow-moving parts like the body and background are classified as low-frequency information.

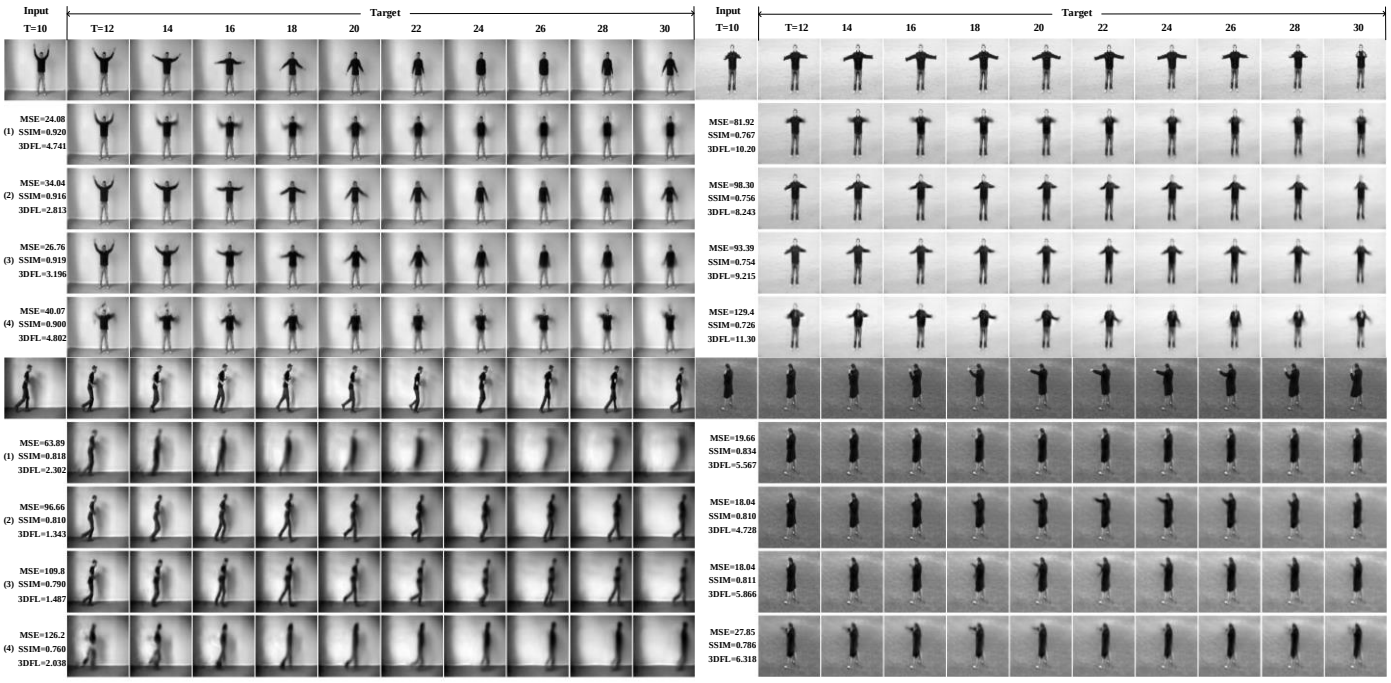

Fig. 2 Metric validation experiments on the KTH dataset. The serial number in parentheses represents that the example in this row is generated by (1) SimVP (2) PredRNNv2 (3) PredRNN (4) ConvLSTM. Three metric performances of each row are labeled on the left. Best viewed by zooming

To incorporate temporal consistency and tackle motion blur, we propose a new metric for video prediction called 3DFL, based on *3D Fourier transform* (3DFFT). 3DFL combines pixel variation information in both spatial and temporal dimensions. The equation for 3DFFT is:

$$F(u,v,w)=\frac{1}{MNT}\sum_{t=0}^{T-1}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y,z)e^{-j2\pi\left(\frac{ux}{M}+\frac{vy}{N}+\frac{tw}{T}\right)} \quad (1)$$

where, M, N, T represent the width, height, and number of frames. $f(x,y,z)$ denotes the pixel value at (x,y,z). We perform 3DFFT on both the predicted frame and the ground truth simultaneously and then calculate their similarity in the frequency domain. Thus, 3DFL is expressed as:

$$3DFL=\sum_{u,v,w}\left\{\log_{10}\left[F(u,v,w)\right]-\log_{10}\left[\hat{F}(u,v,w)\right]\right\}^2 \quad (2)$$

where, $\hat{F}(u,v,w)$ and $F(u,v,w)$ are the 3DFFT of the predicted video and the ground truth, respectively. In practical implementations, 3DFFT can be obtained by performing 3 successive 1DFFT operations on three different dimensions.

B. Metric Validation

We conduct comprehensive validation experiments on the KTH dataset to verify the efficacy of our proposed 3DFL metric against conventional measures like MSE, SSIM, and the qualitative visualizations of predicted frames. Four representative video prediction models are evaluated, including the recurrent network based ConvLSTM, PredRNN, and PredRNNv2, alongside the convolutional model SimVP. These models are tasked with predicting 20 future frames based on 10 previous observations. Fig. 2 illustrates four examples of qualitative visualizations, and the corresponding quantitative performances are labeled on the left.

As illustrated by the 1st Example in Fig. 2, SimVP attains high MSE and SSIM scores but suffers from substantial blurring artifacts in the predicted frames after *t=16*. In contrast, PredRNN and PredRNNv2 generate clearer and more consistent trajectories despite lower MSE and SSIM values. This divergence confirms the deficiencies of traditional metrics in assessing complex video predictions. Only the prediction performance ranking based on our proposed 3DFL metric aligns with the visualization results. We observe similar trends for intricate motions like hand clapping in the 2rd Example, where PredRNN and PredRNNv2 produce perceptually superior predictions compared to SimVP according to 3DFL, despite SimVP achieving higher MSE and SSIM scores. Additional examples further verify that 3DFL rankings consistently match with visual examination across various complex motions, while MSE and SSIM exhibit inconsistencies. This validates the capability of 3DFL as a reliable video prediction quality metric.

V. METHODOLOGY

The qualitative visualization experiments in Section IV revealed two problems with legacy video prediction models. **Firstly,** there was a loss of details in high-frequency information-rich edge regions leading to blurriness. **Secondly,** inaccuracies were observed in predictions containing complex motion information. To address them, we propose a novel video prediction model called SDFNet, which can extract and integrate both spatially encoded shallow and deep-level features, so as to eliminate edge and motion blurring.

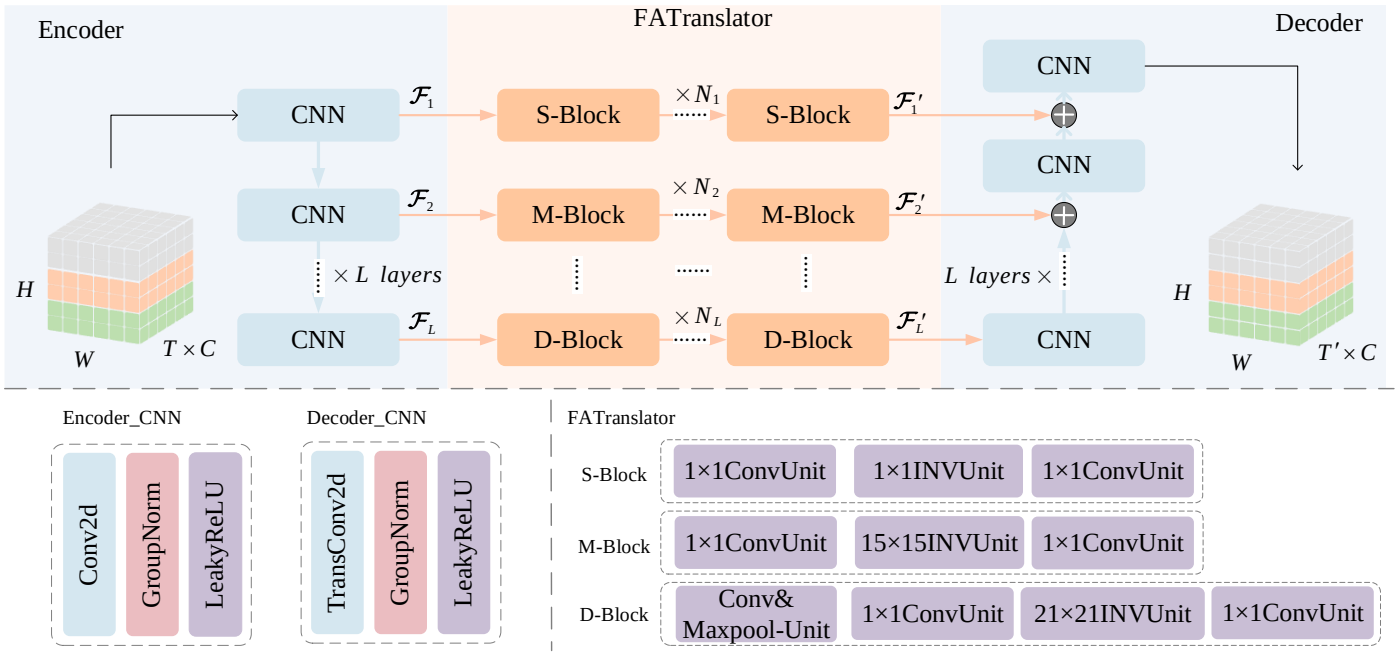

Fig. 3. The overall framework of SDFNet. It contains Encoder, FATranslator and Decoder three fundamental parts. L layers of Convolution and Transpose Conv are stacked to Encoder and Decoder respectively. FATranslator includes three specially designed blocks to the capture of different frequency components.

A. *Problem Formulation*

Video prediction by a DNN model can be formulated as follow. Given an input video sequence $X=\{x_i\},(1\le i\le T)$ of length T frames, we aim to predict the next T' ground truth (GT) frames $Y=\{y_i\},(1\le i\le T')$, where $x_i, y_i \in \mathbb{R}^{C\times H\times W}$ represents the i th frame. C, H, and W represent as a 3D tensor with dimensions of channel, height, and width.

Inputting the sequence X into the mapping function Ψ_{θ} with learnable parameters $\boldsymbol{\theta}$, the DNN model generates the predictive future frames $\hat{Y}=\{\hat{y}_i\},(1\le i\le T')$, $\hat{y}_i \in \mathbb{R}^{C\times H\times W}$. The goal is to minimize the error between the predictive future frames and the GT future frames by learning the optimal parameters $\boldsymbol{\theta}^*$. The optimal parameter $\boldsymbol{\theta}^*$ can be obtained by minimizing the loss function $\mathcal{L}$ in the following formula:

$$\boldsymbol{\theta}^* = \arg\min \mathcal{L}\left\{\Psi_{\theta}(X), Y\right\} \tag{3}$$

B. *SDFNet Model*

Deep learning based video prediction models typically consist of an encoder, translator, and decoder.

- The encoder is a deep neural network with a progressive, multi-level structure, which effectively abstracts and aggregates spatiotemporal information of input frames;
- The translator extracts temporal features from the encoded latent vectors and performs transformations on them;
- The decoder, with a structure similar to that of the encoder, fuses the temporal features processed by the translator and generates predicted video frames.

In legacy video prediction models, the feature is vertically passed from encoder to decoder, and the translator only accepts and processes the features from the last layer of the encoder. However, as discussed and analyzed above, using only the last layer of encoded features is insufficient to solve the edge and motion blurring problem. Based on the inference that deeper networks are better at fitting low spatial frequency information, while high spatial frequency information tends to align with shallow features, we designed a novel video prediction model called SDFNet. This model can extract and integrate spatially encoded features from both shallow and deep levels. SDFNet model is composed of Encoder, FATranslator (cf. Section V-D), and Decoder, as shown in Fig. 3. In SDFNet, features are not only passed vertically for feature extraction to deeper layers but also horizontally input into the FATranslator for temporal feature extraction. The FATranslator accepts features from different depths and uses specially designed blocks to fit different spatial frequency information. The Decoder receives features transformed by different layers in the FATranslator and performs feature fusion based on those transformed features.

To facilitate the introduction of the model, some symbols are defined below. Assume that the model has L layers. $\mathcal{F}_i \in \mathbb{R}^{T\times C'\times H_i'\times W_i'}$, $(1\le i\le L)$ represents the ith feature generated by ith layer of the encoder, where C', H_i' and W_i' are the channel number, height, and width of the encoded sequence, respectively. After being processed by the FATranslator, $\mathcal{F}_i$ turns into $\mathcal{F}_i' \in \mathbb{R}^{T\times C'\times H_i'\times W_i'}$, $(1\le i\le L)$ that serves as the input of the decoder. N_i, $(1\le i\le L)$ represents the number of blocks in the ith layer of the FATranslator. $\mathcal{E}_i, \mathcal{T}_i, \mathcal{D}_i, (1\le i\le L)$ represent the mapping functions of the ith layer in the encoder, FATranslator, and decoder, respectively. Those symbols are summarized in Table I.

The video frames are first fed into the Encoder for

processing. The Encoder consists of L layers, each of which generates a feature $\mathcal{F}_i$. These features are not only vertically input to deeper encoder layers for encoding, but also horizontally fed into the translator, which can be described the following formular:

$$\begin{cases} \mathcal{F}_1 = \mathcal{E}_1(X), i = 1 \\ \mathcal{F}_i = \mathcal{E}_i(\mathcal{F}_{i-1}), i \geq 2 \end{cases} \tag{4}$$

Each $\mathcal{F}_i$ generated by the Encoder will be sent to the FATranslator, where each layer is specially designed to adapt to different frequency information. After being processed by the FATranslator, we have:

$$\mathcal{F}_i' = \mathcal{T}_i(\mathcal{F}_i) \tag{5}$$

Then, $\mathcal{F}_i'$ is entered into the Decoder. The features flow vertically in the Decoder, which can be described as an iteration. Specifically, the feature of the ith layer $\mathcal{F}_i'$, decoded by the corresponding layer of the Decoder, is merges with $\mathcal{F}_{i-1}'$ as the input of the $(i-1)$th layer in the Decoder. Analogously, the rest of the layers do the same, which can be expressed as follows:

$$\begin{cases} \mathcal{F}_i' = \mathcal{D}_i(\mathcal{F}_i'), & i = L \\ \mathcal{F}_i' = \mathcal{D}_i(\mathcal{F}_{i+1}') + \mathcal{F}_i', & 1 \leq i \leq L-1 \end{cases} \tag{6}$$

The following are the implementation details of the Encoder, FATranslator, and Decoder.

TABLE I
SYMBOLS AND THEIR DESCRIPTIONS.

Symbols	Description
L	The number of model layers
i	Count of Layers ($1 \leq i \leq L$)
C'	Hide channel number of the encoded video
H_i'	Height of the encoded video
W_i'	Width of the encoded video
$\mathcal{F}_i$	The ith feature generated by ith layer of the encoder. $\mathcal{F}_i \in \mathbb{R}^{T\times C'\times H_i'\times W_i'}$
$\mathcal{F}_i'$	The ith feature after processed by ith layer of the FATranslator. $\mathcal{F}_i' \in \mathbb{R}^{T\times C'\times H_i'\times W_i'}$
$\mathcal{E}_i$	The mapping functions of the ith layer in the Encoder
$\mathcal{T}_i$	The mapping functions of the ith layer in the FATranslator
$\mathcal{D}_i$	The mapping functions of the ith layer in the Decoder
N_i	The number of blocks in the ith layer of the FATranslator

C. Encoder

The Encoder is responsible for encoding high-dimensional and complex video frame sequences into a lower-dimensional feature representation. This allows the original frame sequence to be mapped to a latent space, where it can be more efficiently processed for video prediction tasks. In our proposed model, the Encoder consists of stacked layers of CNNs. Each CNN layer comprises 3x3 convolution operator, *LayerNorm*, and *LeakyReLU* activation functions. The mapping functions of the Encoder can be represented as follows:

$$\mathcal{E}_i(\cdot) = Leaky\,\mathrm{Re}\,LU\left(LayerNorm\left(Conv(\cdot)\right)\right) \tag{7}$$

Then, each output feature $\mathcal{F}_i$ of the CNN layers in the Encoder is horizontally fed into the FATranslator to extract temporal correlations from the video frame sequence. This horizontal input approach enables diverse depth features to receive corresponding processing within the FATranslator.

D. FATranslator

Typically, a translator learns from the encoded features and captures the time dependency in videos. It transforms the latent features of the observed frames into the latent representations of the predicted frames. However, legacy translators only accept the deep features generated by the last layer of the encoder, neglecting the shallow features from other layers. As a result, their ability to extract high spatial frequency information is weakened. To address this, we propose the Frequency-Adaptive Translator (FATranslator), which has the same number of layers as the Encoder, with each layer corresponding to the respective layer in the Encoder. This enables the separate processing of features generated by different encoder layers.

To adapt the temporal correlation extraction from these spatial features, we design three kinds of blocks in the FATranslator using the involution operator: *S-block*, *M-block*, and *D-block*. These blocks have different structures and kernel sizes, and are specifically designed based on frequency characteristics. Fig. 4 illustrates the diagram of these blocks. The involution operator exhibits channel anisotropy and incorporates adaptive convolution kernels, distinguishing it from conventional convolutions. This unique design enables it

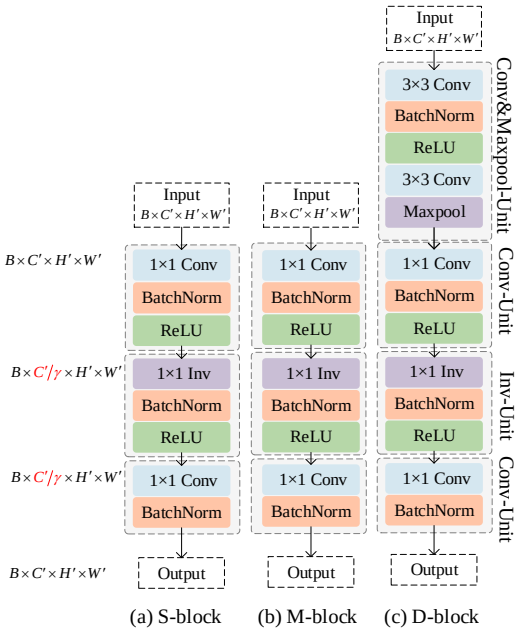

Fig. 4. The diagram of different blocks in Translator. (a) The S-lock. (b) The M-block. (c) The D-block.

to effectively capture temporal dependencies based on spatial features of different depths. The involution operator is well-suited for extracting inter-frame temporal information while still preserving rapid inference speed.

S-block is specially designed for extracting spatially encoded shallow features. As shallow features are well-suited to fit high spatial frequency information, S-block is tailored to its frequency characteristics. It employs 1×1 convolution, 1×1 involution, and 1×1 convolution operators sequentially. The 1 ×1 convolution operator implements channel reduction with rate $\gamma = \{1, 2, 4\}$ and further compresses the features to reduce the number of parameters and computational cost. The 1×1 involution kernel reduces the receptive field, enabling this block to concentrate more on intricate details and high spatial frequency components.

M-block is designed to learn mid-level feature representations that lie between shallow and deep levels. M-block is similar to S-block but has a larger involution kernel to expand the receptive field and capture more contextual and low spatial frequency information.

D-block incorporates a *Conv&Maxpooling* unit to enhance the extraction of low spatial frequency information. The involution kernel of D-block is further expanded to 21, providing the largest receptive field among the three blocks and enabling it to capture the most contextual and low spatial frequency information.

The S-block, M-block, and D-block constitute $\mathcal{T}_1$, $\mathcal{T}_i$, $(2 \le i \le L-1)$, and $\mathcal{T}_L$ respectively. The mapping functions of the FATranslator can be represented as follows:

$$\mathcal{T}_i(\cdot) = \begin{cases} S-block(\cdot), & i = 1 \\ M-block(\cdot), & 2 \le i \le L-1 \\ D-block(\cdot), & i = L \end{cases} \quad (8)$$

E. Decoder

The Decoder is responsible for mapping the processed lower-dimensional features back to predicted video frame sequences. In our proposed model, the Decoder requires L layers of CNNs to receive and decode L different depth features for multi-scale feature fusion. Similar to the Encoder, each layer of the Decoder consists of 3×3 *transpose convolution*, *LayerNorm* and *LeakyReLU*. The mapping functions of the Decoder can be described as:

$$\mathcal{D}_i(\cdot) = Leaky\,\mathrm{Re}LU\left(LayerNorm\left(TransConv(\cdot)\right)\right) \quad (9)$$

Upon the FATranslator extracting inter-frame temporal features, the Decoder fuses these processed *L* layer features to generate predicted video frame sequences.

VI. Experiment

A. Datasets

In this section, we present experiments on various datasets widely used in video prediction to demonstrate the effectiveness of our proposed SDFNet model from three aspects: standard spatiotemporal prediction, variable-length prediction, and RGB video prediction. In the standard spatiotemporal prediction, we selected the Moving MNIST [42] dataset, which is one of the most widely utilized datasets. For variable-length prediction, we conducted experiments on the KTH [41] dataset, which is often used for variable-length prediction in many previous works. For RGB video prediction, we chose the widely used Human3.6 M [43] dataset, which captures complex human actions in the form of RGB images. The details of the three datasets selected are as follows:

- *Moving MNIST*: A synthetically generated dataset containing two digits moving at a constant speed and bouncing inside a 64×64 frame. Each handwritten digit is randomly sampled from the MNIST dataset [44]. Different initial locations and velocities are assigned to each digit, allowing the generation of an infinite number of frame sequences of length 20. Models are trained to predict the future 10 frames based on observing the previous 10 frames.
- *KTH*: A human action dataset consisting of 600 grayscale videos of 15-20 seconds, with a resolution of 120×160 pixels and frame rate of 25 frames per second (fps). It contains 25 individuals performing 6 types of actions, including walking, jogging, running, boxing, hand waving, and hand clapping. Models are trained to generate the subsequent 20 or 40 frames from the last 10 observations.
- *Human3.6M*: A large dataset of human poses with 3.6 million samples, depicting various activities such as taking photos, talking on the phone, posing, greeting, and eating, among others. For experiments, we focus on the *walking* scenario and train the models to predict the following 4 frames given the previous 4 RGB frames.

B. Experiment results

1) Moving MNIST

We evaluate our proposed SDFNet model on the Moving MNIST dataset, and compare it with six RNN-based baselines, including ConvLSTM, PredRNN, PredRNNv2, E3D-LSTM, PhyDnet, and CrevNet. Additionally, we also include the pure CNN-based SimVP model for comparison. All models are reproduced by using their official codes.

SDFNet model is trained on this dataset for 2000 epochs with a batch size of 16, using the Adam optimizer with the OneCycle learning rate scheduler and a learning rate of 0.01. Hide channel $C' = 64$, reduction rate $\gamma = 2$, layer number $L = 4$, the number of blocks in each layer are as follows: {1, 8, 2, 8}. For quantitative evaluation, we use MSE, MAE, SSIM, and our proposed 3DFL metric, which serves as the primary performance ranking criterion. To assess computational efficiency, we report both the training and inference time on a single NVIDIA GeForce GTX 1080Ti GPU.

Table II shows the experimental results. Our proposed SDFNet model performs the best on traditional metrics such as MSE, MAE, and SSIM. Particular, SDFNet outperforms the SOTA model by 26.7% on the MSE metric. Additionally, SDFNet shows significant improvement over the SOTA model on the 3DFL metric. In terms of computational efficiency, our

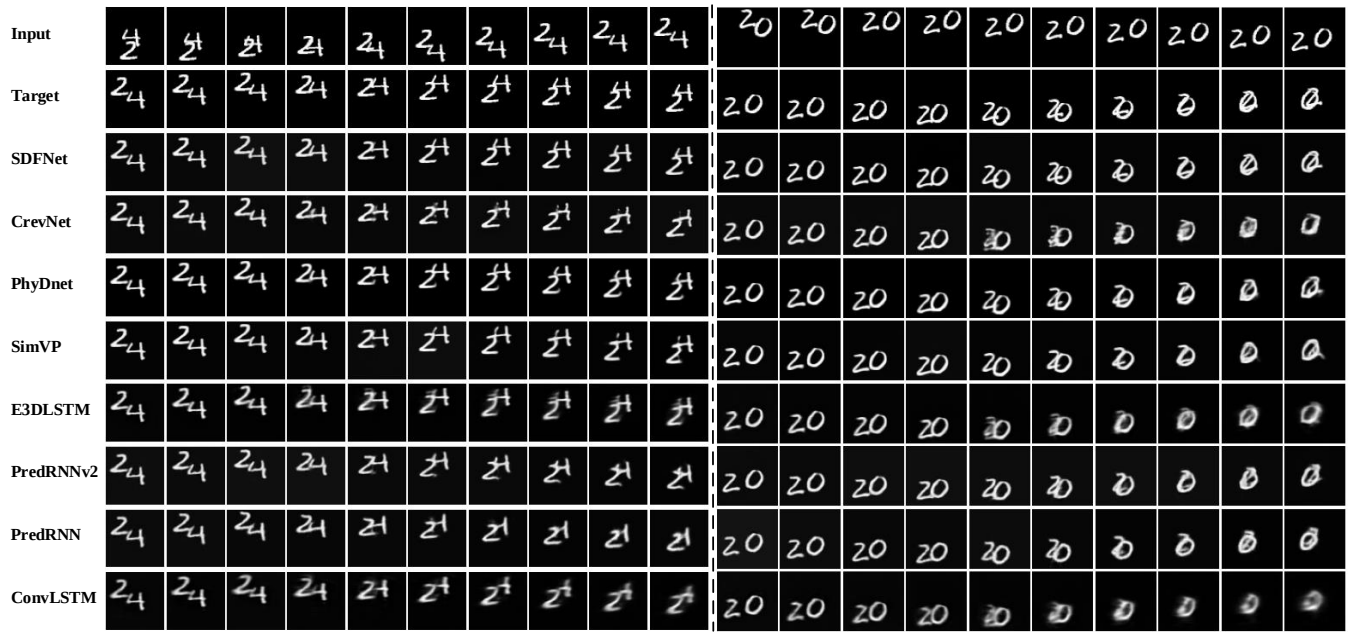

Fig. 5. Two examples of predicted results on the Moving MNIST dataset dataset (10 → 10 frames).

TABLE II
QUANTITATIVE RESULTS OF DIFFERENT MODELS ON THE MOVING MNIST DATASET (10 → 10 FRAMES).

	Training time≈(s) ↓	Inference time≈(ms) ↓	3DFL ↓	MSE ↓	MAE ↓	SSIM ↑
ConvLSTM	797	76.19	8.115	103.3	182.9	0.707
PredRNN	2009	146.28	5.137	56.8	126.1	0.867
PredRNNv2	2104	155.94	4.560	46.5	106.8	0.898
E3D-LSTM	3451	148.55	5.246	41.3	86.4	0.910
PhyDNet	204	72.13	3.500	24.4	70.3	0.947
SimVP	**185**	**17.76**	3.420	23.8	68.9	0.948
Crevnet	922	493.08	3.318	22.3	-	0.949
SDFNet	286	18.91	**2.711**	**17.6**	**57.7**	**0.961**

model's training and inference times are much smaller than those of RNN models and comparable to the highest-performing SimVP model. Thus, SDFNet significantly improves performance while maintaining computational efficiency, achieving the best on various metrics.

Fig. 5 provides two qualitative visualization examples comparing SDFNet with the seven baseline models. The results demonstrate that SDFNet predicts much clearer frames, especially for longer-range predictions. For instance, in the first example, only SDFNet and PhyDnet generate the last prediction frame that accurately recognizes the digits '2' and '4', while other models fail to reconstruct the original digit '4' accurately. In the second example, SDFNet also exhibits better fidelity and temporal consistency in predicting the motion trajectories of digits '0' and '2'. These visualizations provide further evidence of the effectiveness of SDFNet in reducing edge and motion blurring, resulting in improved video prediction accuracy.

To demonstrate that SDFNet effectively captures high spatial frequency information, we transform the predicted results into the frequency domain using a 2DFFT and visualize their frequency spectrum. As is shown in Fig. 6, the low frequency components locate in the center of the spectrum and high frequency components are at the edge. By comparing the spectrums of the predicted results with the ground truth, we observe that SDFNet's predictions contain more high frequency information than baseline models. This supports the notion that SDFNet has a stronger capability at learning high spatial frequency information, thus validating its effectiveness.

2) **KTH**

We experiment on the KTH dataset to investigate SDFNet's ability to predict frames with flexible lengths. This dataset contains 25 individuals performing 6 types of actions. 12 representative baseline models are chosen for comparison, including ConvLSTM, DFN [45], SV2P [46], SAVP-VAE [47], PredRNN, PredRNNv2, MSNET, E3d-LSTM, STMFANet, and SimVP. We reproduce ConvLSTM, PredRNN, PredRNNv2, and SimVP by using their official codes to obtain quantitative performance and qualitative visualization. They are trained with 10 frames to predict 20 or 40 frames. The quantitative performance of the remaining models was based on the data reported in [10].

SDFNet is trained on this dataset for 200 epochs with a batch size of 4, using the Adam optimizer with the OneCycle learning rate scheduler and a learning rate of 0.01. Hide channel $C' = 64$, reduction rate $\gamma = 4$, layer number $L = 4$,

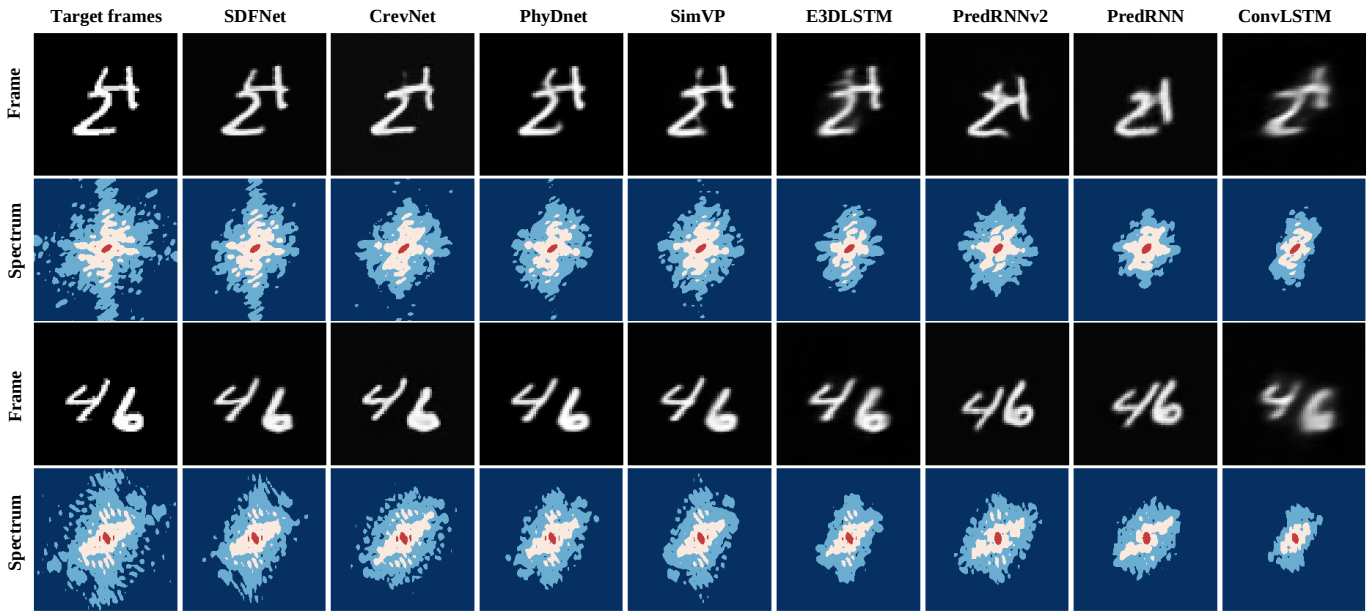

Fig. 6. The spectrum comparison and original image comparison between ground truth and predicted frames generated by SDFNet and other seven chosen baselines. Rows 1and 3 are the last original frames in predicted videos, 2 and 4 are their spectrum.

TABLE III
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE KTH DATASET (10 → 20 FRAMES & 10 → 40 FRAMES).

Methods	(KTH)10→20				(KTH)10→40			
	3DFL ↓	MSE ↓	SSIM↑	PSNR↑	3DFL ↓	MSE ↓	SSIM ↑	PSRN↑
DFN	-	-	0.794	27.26	-	-	0.652	23.01
SV2Pi	-	-	0.826	27.56	-	-	0.778	25.92
SV2Pv	-	-	0.838	27.79	-	-	0.789	26.12
SVAP-VAE	-	-	0.852	27.77	-	-	0.811	26.18
E3d-LSTM	-	-	0.879	29.31	-	-	0.810	27.24
STMFANet	-	-	0.893	29.85	-	-	0.851	27.56
ConvLSTM	7.779	63.08	0.712	23.56	13.636	92.54	0.639	22.85
PredRNN	5.953	38.75	0.839	27.55	12.39	91.70	0.703	24.16
PredRNNv2	5.311	39.11	0.865	28.47	9.058	59.13	0.741	25.21
SimVP	7.098	43.45	0.905	33.72	10.49	51.86	0.886	32.93
SDFNet	**4.598**	45.82	**0.912**	**34.18**	**6.25**	54.31	**0.891**	**33.17**

the number of blocks in each layer are as follows: {1, 5, 2, 6}. For quantitative evaluation, we use MSE, MAE, SSIM, and our proposed 3DFL metrics. 3DFL serves as the primary performance ranking criterion.

Table III presents the quantitative comparison of SDFNet and other chosen models on the KTH dataset. SDFNet achieves the best performance on the 3DFL metric. In addition, SDFNet outperforms the SOTA model in terms of SSIM metric and shows improvements of 1.36% and 0.73% in PSNR for (10 → 20) and (10 → 40) settings, respectively. It's worth noting that the rankings based on the 3DFL metric are considered more reasonable and reliable than the MSE and SSIM metrics, as demonstrated in Section IV.

Fig. 7 provides three visualization comparison examples for SDFNet and the four baseline models reproduced by using their official code.

In the first example, the video depicts a person jogging. The predictions of SimVP and ConvLSTM quickly become blurred in the first few frames, resulting in the person becoming an indistinguishable black shadow. SDFNet, PredRNN, and PredRNNv2 are able to generate higher-quality predictions, but their clarity varies. Notably, SDFNet preserves fine details, such as the person's calf that has not yet disappeared, while PredRNN and PredRNNv2 fail to maintain such details.

The second example involves a person performing a waving motion, making it difficult to predict the movement of the arm. SimVP and ConvLSTM only correctly predict the stationary body and background, which are easier to be predicted and contribute significantly to the MSE and SSIM metrics. However, SDFNet, PredRNN, and PredRNNv2 can predict and generate the arm's movement. Particularly, SDFNet produces frames with less blur and higher clarity. In the final frame, the person's arm is raised in a V-shape. SDFNet precisely predicts this motion with high clarity, while PredRNN and PredRNNv2 predict the arm in a raised position with blurred forearms.

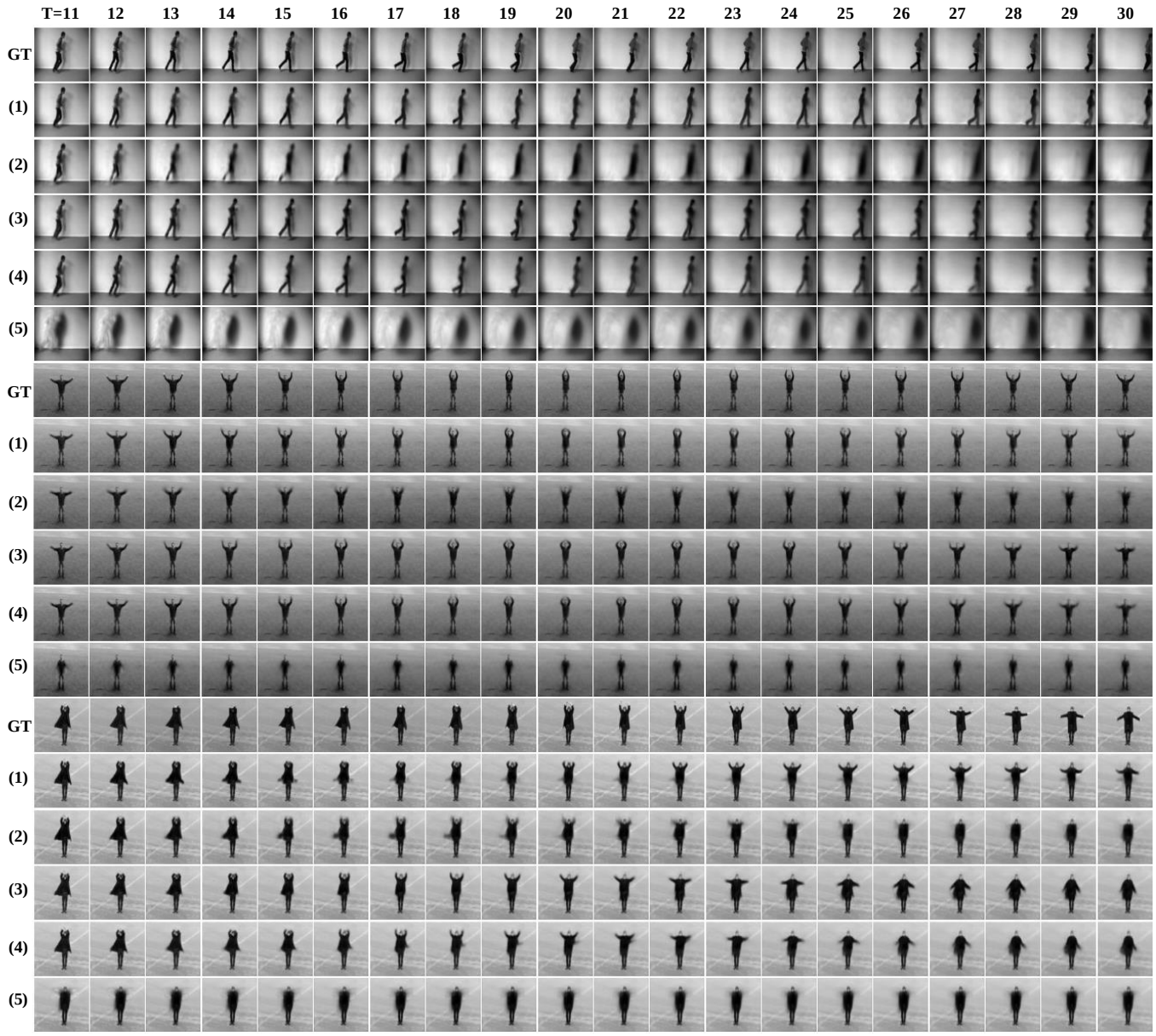

Fig. 7. Three examples of predicted results on the KTH dataset (10 → 20 frames). (1) SDFNet (2) SimVP (3) PredRNNv2 (4) PredRNN (5) ConvLSTM.

In the third example, the person performs a waving motion, but the background differs from the second scenario. SDFNet remains the best performer, followed by PredRNNv2, PredRNN, SimVP, and ConvLSTM. In the last frame, the person keeps his arm raised flat. Only SDFNet can accurately distinguish the raised flat arm, while the other models fail to do so. Moreover, the frame-by-frame observation demonstrates that SDFNet performs well in predicting motion trajectories with superior temporal consistency.

In general, SDFNet effectively reduces edge and motion blurring caused by missing high-frequency components and achieves significant improvements in fidelity and temporal consistency, which demonstrates its validity.

3) **Human3.6M**

Human 3.6M is another human pose dataset similar to KTH but with more complicated backgrounds, thus posing difficulties for video prediction. For our experiments, we focused on the *walking* scenario. We select ConvLSTM, PredRNN, PredRNNv2, and SimVP as the baseline models for reproduction and comparison. All models are trained with 4 previous frames to predict the following 4 frames. We trained our SDFNet model for 100 epochs using a batch size of 8 and the Adam optimizer with a learning rate of 0.01. Hide channel $C' = 64$, reduction rate $\gamma = 4$, layer number $L = 4$, the number of blocks in each layer are as follows: {1, 5, 2, 6}. For the quantitative evaluation, we utilized MSE, MAE, SSIM, and our proposed 3DFL metrics, with 3DFL serving as the primary performance ranking criterion.

Table IV the quantitative performance comparison of SDFNet and the baselines on the Human3.6M dataset. Although SDFNet does not show improvement in MSE, it achieves the best performance on the 3DFL and SSIM metrics. In particular, the 3DFL metric demonstrated a 7.94%

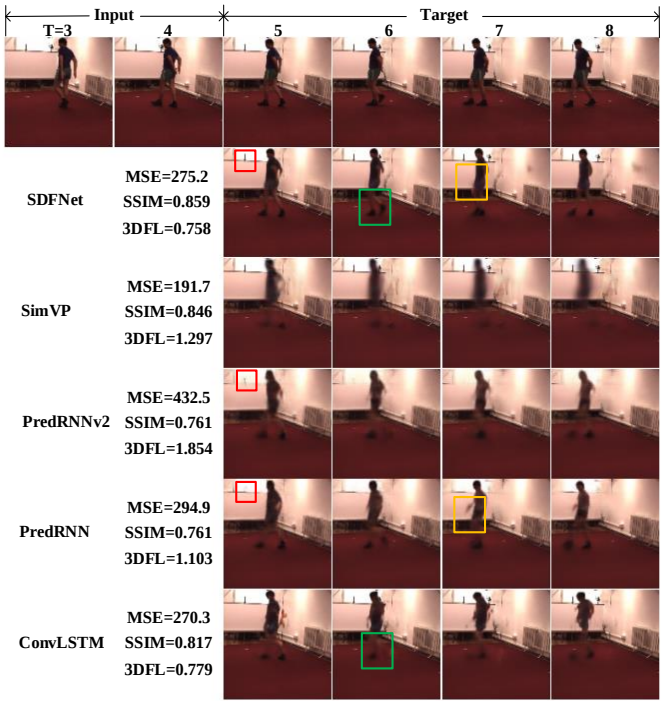

Fig. 8. Qualitative comparison on the Human3.6M (4 →4 frames). The first row is the input (only visualized last two frames) and target.

TABLE IV
QUANTITATIVE RESULTS OF DIFFERENT METHODS ON THE HUMAN 3.6M DATASET (4 → 4 RGB FRAMES).

	3DFL ↓	MSE/10 ↓	MAE-	SSIM-
MIM	-	42.9	17.8	0.790
E3D-LSTM	-	46.4	16.6	0.869
PhyDnet	-	36.9	16.2	0.901
ConvLSTM	2.815	50.4	18.9	0.776
PredRNN	3.161	47.8	18.7	0.753
PredRNN++	3.597	47.5	18.2	0.771
SimVP	3.676	31.6	15.1	0.904
SDFNet	**2.693**	32.6	15.3	**0.910**

improvement, confirming the effectiveness of SDFNet.

Fig. 8 shows the visualization comparison of the predictions made by SDFNet and the four baseline models reproduced with official code. The observations are as follows:

- As illustrated in the red boxes, PredRNN and PredRNNv2 produce inaccurate predictions of the background. SDFNet outperforms them in capturing spatial information, leading to more accurate background predictions.
- The predictions generated by SDFNet exhibit accurate and clear edges around the human, representing a significant improvement in fidelity. As shown in the green boxes, only SDFNet can generate clear edges, while the others suffer from varying degrees of blurring. This is because SDFNet places greater emphasis on high spatial frequency information in complex scenes.
- SDFNet achieves high temporal consistency and accuracy in predicting motion trajectories, surpassing other baseline models. While PredRNN can also correctly capture motion trajectories, it produces incorrect predictions in some details, as highlighted in the yellow boxes. This highlights the effectiveness of the involution operation we introduced for temporal feature extraction.

In conclusion, SDFNet effectively reduces edge and edge blurring caused by missing high spatial frequency components and achieves significant improvements in fidelity and temporal consistency on all the three datasets. These experimental results demonstrate the validity of SDFNet. The performance rankings according to our proposed 3DFL align well with human intuition across various experiments, offering a more reasonable criterion than traditional metrics.

C. *Ablation study*

Our proposed model leverages two key points, namely the shallow and deep features combined architecture (*SDF architecture*) and the FATranslator, where the FATranslator is based on the introduced *Involution* operator. Therefore, we conduct ablation studies on the Moving MNIST dataset to evaluate the effectiveness of the *SDF architecture* and the introduced *Involution* operator. To be specific, we add two variants of SDFNet for experiment, namely the w/o SDF and w/o INV models. The w/o SDF model signifies the utilization of solely the deep spatial feature extracted from the last layer of the encoder. The w/o INV model represents replacing the involution with convolution in the FATranslator.

In the first experiment, we validate the impacts of the SDF architecture and the involution operation on SDFNet's accuracy and computational efficiency. Six baseline models are chosen for comparison, including ConvLSTM, PredRNN, PredRNN++, E3D-LSTM, SimVP, and CrevNet. Table V summarize the evaluated 3DFL, MSE, FLOPs, and #Params metric values for the models. Although the w/o SDF model results in 3.247 on 3DFL and 22.1 on MSE, it still slightly outperforms the current SOTA model while only having 0.79 GFLOPs. Although the w/o INV model surpasses the SOTA model with a 3DFL value of 2.856 and an MSE value of 19.5, this achievement is accompanied by a notable rise in computational complexity and parameter count. Specifically, the computational demands increase substantially, reaching 34.11 GFLOPs, while the number of parameters rises to 603M. In other words, the SDF architecture can significantly enhance video prediction accuracy, while the involution operation not only improves accuracy but also reduces the computational complexity and number of parameters of the model.

TABLE V: THE ABLATION STUDY ON THE MOVING MNIST DATASET SHOWS THE INFLUENCE OF SPATIAL SHALLOW FEATURES AND INVOLUTION.

Models	3DFL	MSE	FLOPs	# Params
ConvLSTM	8.115	103.3	107.4G	-
PredRNN	5.137	56.8	192.9G	-
PredRNN++	4.560	46.5	106.8G	-
E3D-LSTM	5.246	41.3	381.3G	-
SimVP	3.420	23.8	1.676G	29.16M
CrevNet	3.318	22.3	1.633G	-
Ours (w/o SDF)	3.247	22.1	0.79G	13.21M
Ours (w/o INV)	2.856	19.5	34.11G	603.56M
Ours	**2.711**	17.6	1.793G	21.67M

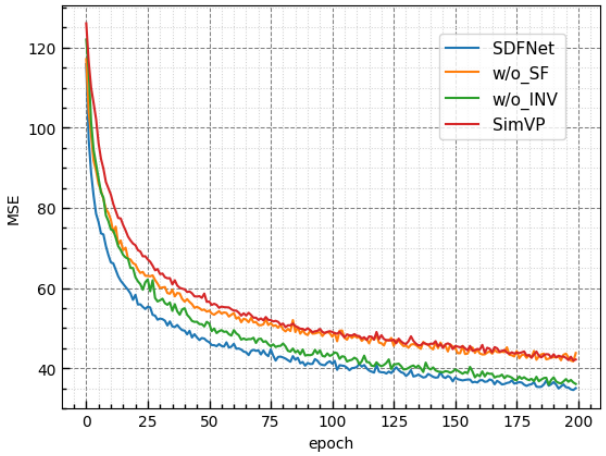

Fig. 9. The ablation study of convergence speed on Moving MNIST dataset. This figure shows the performances of the first 200 epochs in 2k epochs.

In the second experiment, we assess the impact of the SDF architecture and the involution operation on the convergence speed of model training. To provide a comprehensive comparison, we also incorporate SimVP due to the similarity of its *encoder-decoder* structure with SDFNet. In Fig. 9, we depict the comparison of convergence speeds during the initial 200 epochs, within a total training span of 2000 epochs. The blue line corresponds to SDFNet, which integrates both the SDF architecture and the involution operation, showcasing the fastest convergence speed along with superior MSE performance. Representing the w/o INV model, the green line features the SDF architecture alone and displays commendable convergence speed. In contrast, the red and yellow lines, lacking the SDF architecture, exhibit comparatively slower convergence. These findings underscore the effective role of the SDF architecture in accelerating model convergence.

The third experiment illustrates the impact of the SDF architecture on high spatial frequency information within the prediction outcomes. This is visually depicted in Fig. 10, showcasing a comparison. It becomes evident that both SDFNet and the w/o INV model, integrating the SDF architecture, retain a greater amount of high spatial frequency information in their predictions, when compared to the w/o SDF model. This visual analysis furnishes empirical evidence that underscores the SDF architecture's effectiveness in adeptly capturing high spatial frequency information.

In summary, the SDF architecture not only significantly enhances the model's overall performance and convergence speed but also bolsters its ability to capture high spatial frequency information. The incorporation of the involution operation further contributes to performance enhancements, expedited convergence, as well as reduced computational complexity and parameters stemming from extensive convolutional kernels. These findings underscore the prowess of our proposed model in video prediction tasks.

VII. Conclusion

In this paper, we have investigated the issue of edge and motion blurring in video prediction, which arises due to the predictive model's insufficient attention to high spatial frequency information. Firstly, since legacy performance metrics like MSE and SSIM cannot decently evaluate this deficiency, we designed a 3DFL metric to better assess the similarity of predictive video frames. Secondly, aiming to alleviate edge and motion blurring, and motivated by the observation that shallow features inherently possess more high spatial frequency information compared to deep features, we proposed a novel video prediction model called SDFNet which can extract and integrate both spatially encoded shallow and deep-level features. Thirdly, to accommodate SDFNet's multi-branch input structure, we have elaborated a FATranslator, which leverages involution operation to adaptively extract inter-frame temporal dependencies from different spatial encoding layers. Finally, extensive experiments have been conducted to evaluate SDFNet against baseline models. The results consistently demonstrate the superiority of our model in terms of reducing edge and motion blurring and enhancing overall prediction accuracy.

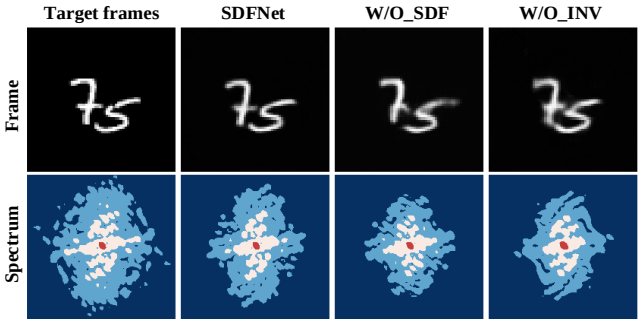

Fig. 10. The ablation study of spectrums on the Moving MNIST dataset shows the influence of the SDF architecture.

References

[1] Y. Wang, J. Zhang, H. Zhu, M. Long, J. Wang and P. S. Yu, "Memory in Memory: A Predictive Neural Network for Learning Higher-Order Non-Stationarity From Spatiotemporal Dynamics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019*, pp. 9146-9154.

[2] Shi, X., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., and Woo, W. C., "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[3] B. Klein, L. Wolf, and Y. Afek, "A dynamic convolutional layer for short rangeweather prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4840–4848

[4] Q. Men, E. S. L. Ho, H. P. H. Shum, and H. Leung, "A Quadruple Diffusion Convolutional Recurrent Network for Human Motion Prediction," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 9, pp. 3417-3432, Sept. 2021, doi: 10.1109/TCSVT.2020.3038145.

[5] Li, Z. Zhou, and X. Liu, "Multi-person pose estimation using bounding box constraint and lstm," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2653–2663, Oct. 2019.

[6] Kong, Y., and Fu, Y., "Human action recognition and prediction: A survey," *International Journal of Computer Vision*, vol. 130, no. 5, pp. 1366-1401, 2022.

[7] Hu, W. Zhan, and M. Tomizuka, "Probabilistic prediction of vehicle semantic intention and motion," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Jun. 2018, pp. 307–313.

[8] Y. LeCun, Y. Bengio, and G. E. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] Xu, J., Ni, B., Li, Z., Cheng, S., and Yang, X., "Structure preserving video prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1460-1469.
[10]Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., and Argyros, A., "A review on deep learning techniques for video prediction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2806-2826, 2020.
[11]Jin, Y. Hu, Q. Tang, J. Niu, Z. Shi, Y. Han, and X. Li, "Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4554–4563.
[12]Z. Liu, R. A. Yeh, X. Tang, Y. Liu, and A. Agarwala, "Video frame synthesis using deep voxel flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4473–4481.
[13]Z. Xu, Y. Wang, M. Long, J. Wang, and M. KLiss, "Predcnn: Predictive learning with cascade convolutions." in *IJCAI*, 2018, pp. 2940–2947.
[14]N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. ICML*, 2015, pp. 843–852.
[15]X. Liang, L. Lee, W. Dai, and E. P. Xing, "Dual motion GAN for future flow embedded video prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1762–1770.
[16]Gao, Z., Tan, C., Wu, L., and Li, S. Z. "SimVP: Simpler Yet Better Video Prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 3170-3180.
[17]J. Revaud, P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "EpicFlow: Edge-preserving interpolation of correspondences for optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1164–1172
[18]S. Li, J. Fang, H. Xu, and J. Xue, "Video frame prediction by deep multi-branch mask network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1283–1295, Apr. 2021.
[19]Lin, X., Zou, Q., Xu, X., Huang, Y., and Tian, Y., "Motion-aware feature enhancement network for video prediction." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 688-700, 2020.
[20]X. Chen, W. Wang, J. Wang, and W. Li, "Learning object-centric transformation for video prediction," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 1503–1512
[21]J. Wang, W. Wang, and W. Gao, "Predicting diverse future frames with local transformation-guided masking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3531–3543, Dec. 2019.
[22]X. Lin, Q. Zou, X. Xu, Y. Huang, and Y. Tian, "Motion-aware feature enhancement network for video prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 688–700, Feb. 2021.
[23]X. Chen and W. Wang, "UNI-and-BI-directional video prediction via learning object-centric transformation," *IEEE Trans. Multimedia*, vol. 22, no. 6, pp. 1591–1604, Jun. 2020.
[24]H.-k. Chiu, E. Adeli, and J. C. Niebles, "Segmenting the future," *IEEE Robotics Autom. Lett.* vol. 5, no. 3, pp. 4202–4209, 2020
[25]Y. Wang, L. Jiang, M.-H. Yang, L.-J. Li, M. Long, and L. Fei-Fei, "Eidetic 3D LSTM: A model for video prediction and beyond," in *Proc. Int. Conf. Learn. Representations Poster*, 2019.
[26]Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600-612, 2004.
[27]Y. Wang, M. Long, J. Wang, Z. Gao, and P. S. Yu, "PredRNN: Recurrent neural networks for predictive learning using spatiotemporal lstms," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 879–888.
[28]Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, "Predrnn++: Towards A resolution of the deep-in-time dilemma in spatiotemporal predictive learning," in *Proc. 35th Int. Conf. Mach. Learn., 2018*, pp. 5123–5132.
[29]Y. Wang, H. Wu, J. Zhang, Z. Gao, J. Wang, P. S. Yu, and M. Long, "Predrnn: A recurrent neural network for spatiotemporal predictive learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2208-2225, 2022.
[30]Guen, Vincent Le, and Nicolas Thome, "Disentangling physical dynamics from unknown factors for unsupervised video prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11474-11484.
[31]W. Yu, Y. Lu, S. Easterbrook, and S. Fidler, "Efficient and information-preserving future frame prediction and beyond," in *International Conference on Learning Representations*, 2019.
[32]Lee, S., Kim, H. G., Choi, D. H., Kim, H. I., and Ro, Y. M., "Video prediction recalling long-term motion context via memory alignment learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3054-3063.
[33]Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P., "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600-612, 2004.
[34]Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O., "The unreasonable effectiveness of deep features as a perceptual metric," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586-595
[35]Li, D., Hu, J., Wang, C., Li, X., She, Q., Zhu, L., and Chen, Q., "Involution: Inverting the inherence of convolution for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12321-12330.
[36]Farazi, H., Nogga, J., and Behnke, S., "Local frequency domain transformer networks for video prediction," in *International Joint Conference on Neural Networks* (IJCNN), July 2021, pp. 1-10.
[37]Wang, H., Wu, X., Huang, Z., and Xing, E. P., "High-frequency component helps explain the generalization of convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8684-8694.
[38]Xu, Z. Q. J., Zhang, Y., Luo, T., Xiao, Y., and Ma, Z., "frequency Principle: Fourier Analysis Sheds Light on Deep Neural Networks," *Communications in Computational Physics*, vol. 28, no. 5, pp. 1746-1767, 2020.
[39]Xu, Z. J., and Zhou, H., "Deep frequency principle towards understanding why deeper learning is faster," In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 12, pp. 10541-10550, May 2021.
[40]Huang, C., Li, J., Li, B., Liu, D., and Lu, Y., "Neural Compression-Based Feature Learning for Video Restoration," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5872-5881.
[41]Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, vol. 3. IEEE, 2004, pp. 32–36.
[42]Srivastava, N., Mansimov, E., and Salakhudinov, R., "Unsuper-vised learning of video representations using lstms," in *International conference on machine learning. PMLR*, 2015, pp. 843-852.
[43]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*, vol. 36, no. 7, pp. 1325–1339, Jun. 2013.
[44]Yann LeCun, Corinna Cortes, and CJ Burges. "MNIST handwritten digit database". In: ATT Labs [Online]. 2 (2010). URL: http://yann.lecun.com/exdb/mnist.
[45]X. Jia, B. De Brabandere, T. Tuytelaars, and L. V. Gool, "Dynamic filter networks," *Advances in neural information processing systems*, vol. 29, 2016.
[46]M. Babaeizadeh, C. Finn, D. Erhan, R. H. Campbell, and S. Levine, "Stochastic variational video prediction," in *6th International Conference on Learning Representations, ICLR 2018*. 2018.
[47]Lee, A. X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., and Levine, S., "Stochastic adversarial video prediction." *arXiv preprint arXiv:1804.01523* (2018).