This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SRVP: Strong Recollection Video Prediction Model Using Attention-Based Spatiotemporal Correlation Fusion

Yuseon Kim[1,2] Kyongseok Park[1,2*]
[1]Korea Institute of Science and Technology Information (KISTI)
[2]Department of Applied AI, University of Science and Technology (UST)
{yskblue,gspark}@kisti.re.kr

Abstract

Video prediction (VP) generates future frames by leveraging spatial representations and temporal context from past frames. Traditional recurrent neural network (RNN)-based models enhance memory cell structures to capture spatiotemporal states over extended durations but suffer from gradual loss of object appearance details. To address this issue, we propose the strong recollection VP (SRVP) model, which integrates standard attention (SA) and reinforced feature attention (RFA) modules. Both modules employ scaled dot-product attention to extract temporal context and spatial correlations, which are then fused to enhance spatiotemporal representations. Experiments on three benchmark datasets demonstrate that SRVP mitigates image quality degradation in RNN-based models while achieving predictive performance comparable to RNN-free architectures.

1. Introduction

Video prediction (VP) involves generating future video sequences by learning spatial correlations and temporal information from past frames; it is also considered a spatiotemporal forecasting problem. Unlike single-image prediction, VP requires estimating temporal dynamics and spatial representations simultaneously, making it a complex and challenging task.

Traditional approaches for processing image-sequence datasets, such as videos, rely on recurrent neural network (RNN)-based models such as long short-term memory (LSTM) networks [9, 11], which excel in time series forecasting and have applications beyond VP. Early VP models used a simple LSTM-based encoder-decoder framework [20], where the encoder compresses input into a lower-dimensional hidden representation, and the decoder predicts future frames. Subsequently, convolutional LSTM (ConvLSTM) [19] extended LSTM's input-to-state and state-to-state transitions into a convolutional structure.

Building on ConvLSTM, various VP models have been proposed to enhance long-term video dependency capture [4, 10, 13, 21, 25–29]. These methods include augmenting ConvLSTM cells with spatiotemporal states and deepening state circulation through zigzag transitions [25], extending dual-memory units into a cascaded structure along the time axis [26], modeling non-stationary and stationary spatiotemporal dynamics [28], generalizing ConvLSTM to higher dimensions [21], integrating transient and motion trends in video sequences [29], and incorporating self-attention into ConvLSTM cells [13].

Although these RNN-based methods can naturally model the temporal transitions of video, they can be highly sensitive to object motion. Spatiotemporal memory cells store information from previous time steps and update the spatial information of the next time step accordingly. Owing to this inherent characteristic, the model tends to predict average values in regions of change, minimizing error but leading to blurring in generated frames. Moreover, errors accumulate over time, further degrading accuracy in long-term predictions.

In addition, vision transformer (ViT)-based methods have been explored for video analysis [2, 3, 14, 15, 31]. However, most focus on extracting high-dimensional image features for tasks such as classification, with limited attempts at VP. ViT-based VP approaches [16, 30] aim to mitigate error accumulation in long-term predictions by generating future frames in a single step. However, unlike RNN models that employ gating mechanisms to regulate temporal variations, ViTs lack an explicit mechanism for controlling temporal changes. Therefore, additional techniques such as positional encoding and 3D convolutions are required for auto-regression, but their effectiveness in capturing both short- and long-term dependencies remains unclear.

Moreover, input image size significantly impacts ViT-based VP performance. As ViTs partition images into non-overlapping patches, larger images force each patch to cover

*Corresponding author

a broader area, reducing fine detail capture and potentially decreasing accuracy. In contrast, smaller patches increase the number of boundaries within an image, disrupting spatial continuity. In addition, a larger number of patches also raises computational costs and may require extensive hyperparameter tuning to identify the optimal patch size.

Therefore, we propose the strong recollection VP (SRVP) model, which integrates an attention-based approach to improve the retention of spatially varying object representations over longer durations while preserving the ability of RNNs to estimate temporal context. SRVP enhances long-term memory of moving objects by referencing spatial features from past frames through temporal attention and capturing spatial correlations between pixels via spatial self-attention.

2. Background

The ConvLSTM unit enables spatiotemporal modeling through three interacting gates: input, forget, and output [21]. However, because ConvLSTM merely extends conventional LSTM gate operations into a convolutional structure, it inherently prioritizes temporal variations.

To address this limitation and enhance the extraction of spatiotemporal representations, spatiotemporal LSTM (ST-LSTM) [25] was introduced. The ST-LSTM unit employs a dual-memory structure, integrating a spatiotemporal memory cell alongside the conventional ConvLSTM memory cells. This design allows the model to capture both low-dimensional image features and high-level abstract representations by updating spatiotemporal memory cell states in a zigzag pattern. However, this complex nonlinear transition structure is prone to the vanishing gradient problem, where gradient magnitudes decay exponentially during backpropagation through time (BPTT). To mitigate this, causal LSTM and the gradient highway unit (GHU) were proposed [26]. Causal LSTM extends the dual-memory structure of ST-LSTM into a cascaded architecture, improving structured transitions across deeper temporal stages, expanding the receptive field, and enhancing long-term retention of object features. In addition, the GHU shortens backpropagation paths from distant inputs to future outputs, reducing gradient decay. Subsequently, the memory in memory (MIM) [28] block was introduced to capture complex spatiotemporal variations. Using a two-cascaded, self-renewing memory module, MIM leverages difference signals between adjacent recurrent states to capture both stationary and non-stationary spatiotemporal dynamics. A network with stacked MIM blocks enabled the modeling of high-order non-stationary phenomena in complex systems such as traffic flow and weather patterns.

The gating mechanism in these RNN-based approaches sophisticatedly regulates information retention and forgetting, allowing for the effective learning of natural temporal information. These models update the spatiotemporal state for the next frame based on information from prior time steps. However, in regions with continuous motion—for instance, with a walking pedestrian or a moving vehicle—repeated updates degrade spatial information retention, weakening object appearance recall over time.

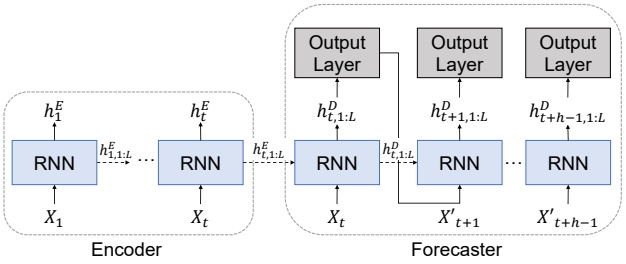


Figure 1. Baseline VP model. The encoder network includes L-RNN layers, whereas the forecaster network comprises L-RNN layers and a CNN-based output layer.

Consequently, the model tends to predict average values in dynamically changing areas to minimize error, leading to more pronounced blurring effects in motion regions compared to static ones. As VP tasks require not only high quantitative performance but also visually coherent frame generation, reducing blurring artifacts remains a critical challenge.

Accurately preserving the intrinsic representation of moving objects at each time step is key to generating sharper and more natural video sequences. The proposed method follows this principle by introducing two attention modules designed to preserve spatially varying object representations while incorporating the temporal context retention capability of RNNs. The subsequent section details our approach.

3. Methods

3.1. Baseline Architecture

We adopted an RNN-based encoder-forecaster architecture [5, 22] as the baseline VP model, as depicted in Fig. 1. This architecture comprises encoding and forecasting components, resembling the ConvLSTM encoder-decoder model [19]. The encoding network comprises L-RNN layers, whereas the forecasting network comprises L-RNN layers and a convolutional neural network (CNN)-based output layer. The encoding network sequentially processes N historical frames, $X_{1:t} \in \mathbb{R}^{N\times C\times H\times W}$, as input, learning the video dynamics over time. Each image frame is encoded layer by layer, progressively transforming it into $h^E_{1:t} \in \mathbb{R}^{N\times M\times H\times W}$, which captures high-dimensional spatiotemporal features. The hidden states of each layer are propagated along the temporal axis (horizontally), allowing for simultaneous updates of both temporal context and spa-

tial representations. In the forecasting network, the final hidden state of the encoder, $h^E_{t,1:L} \in \mathbb{R}^{L\times M\times H\times W}$, serves as the initial state for predicting future frame sequences.

In this study, the number of layers (L) was four, and the size of the hidden states (M) in each layer was 128 in both the encoder and the forecaster.

3.2. Convolutional Gated Recurrent Unit

This study used the convolutional gated recurrent unit (ConvGRU) as the RNN layer of SRVP. Similar to ConvLSTM, ConvGRU is a spatiotemporal recurrent unit that extends the conventional GRU [5, 7] by incorporating spatial dimensions. It comprises two principal gating mechanisms: reset gate r_t and update gate z_t, which are formulated as follows:

$$\begin{aligned} r_t &= \sigma(W_{xr} * X_t + W_{hr} * h_{t-1} + b_r) \\ g_t &= \tanh(W_{xg} * X_t + W_{hg} * (r_t \circ h_{t-1}) + b_g) \\ z_t &= \sigma(W_{xz} * X_t + W_{hz} * h_{t-1} + b_z) \\ h_t &= (1 - z_t) \circ g_t + z_t \circ h_{t-1} \end{aligned} \tag{1}$$

where σ represents the sigmoid function, and $*$ and $\circ$ denote the convolution operator and Hadamard product, respectively. The reset gate determines how much of the past information should be forgotten by considering both the current input and historical data, while the update gate regulates the extent to which newly computed information, along with past information, is incorporated into the current state.

The ConvGRU unit learns spatiotemporal representations through a more simplified hidden state computation than ConvLSTM. This streamlined architecture improves the responsiveness and efficiency of the model in capturing temporal variations [7].

Fig. 2 illustrates the overall architecture of the SRVP model. When the prediction target is X_{t+1}, the ConvGRU layers in the encoder take the historical frames, $X_{1:t}$, as input and generate reference features $h^E_{1:t}$. The ConvGRU layers in the forecaster receive the current frame, X_t, and the final hidden state of the encoder as initial states, producing the target features $h^D_{t,1:L} \in \mathbb{R}^{L\times M\times H\times W}$.

3.3. Spatiotemporal Correlation Fusion

To effectively predict both short- and long-term variations in video sequences, a spatiotemporal prediction model must fulfill the following two key requirements:

1. The spatiotemporal memory should concentrate on capturing the temporal context of the video (e.g., the position of objects in the next frame) by appropriately storing and updating information.
2. The model should accurately recall the intrinsic representation of objects (e.g., the structure of a human body or the shape of digits).

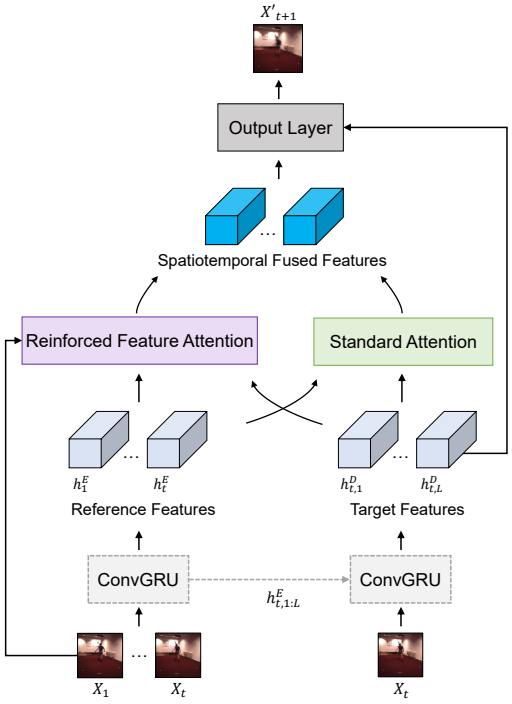


Figure 2. Overall SRVP architecture, primarily comprising two modules, standard attention (SA) and reinforced feature attention (RFA), which together extract spatiotemporal fused features.

To address this challenging task, we propose a method that leverages the attention mechanism to extract various spatiotemporal features from video sequences. The proposed approach initially focuses separately on temporal context estimation and spatial correlation enhancement and then integrates these two distinct feature representations to generate spatiotemporal features, which are used to generate subsequent frames.

As shown in Fig. 2, SRVP extends the baseline architecture by incorporating attention modules. It comprises two main modules, standard attention (SA) and reinforced feature attention (RFA), which together extract integrated spatiotemporal features.

The SA module comprises temporal attention, spatial self-attention, and cross-attention (Fig. 3). In machine translation, the attention mechanism identifies the most critical elements within an input sentence that contribute to understanding the overall context [1]. From the VP perspective, this can be analogous to identifying the most relevant frames from the historical sequences that best inform the current prediction. Furthermore, the self-attention mechanism [23] is formulated as a process for exploring spatial correlations between pixels in the spatiotemporal features.

The recurrent memory outputs, $h^E_{1:t}$ and $h^D_{t,1:L}$, contain M hidden states per pixel, encoding both temporal correla-

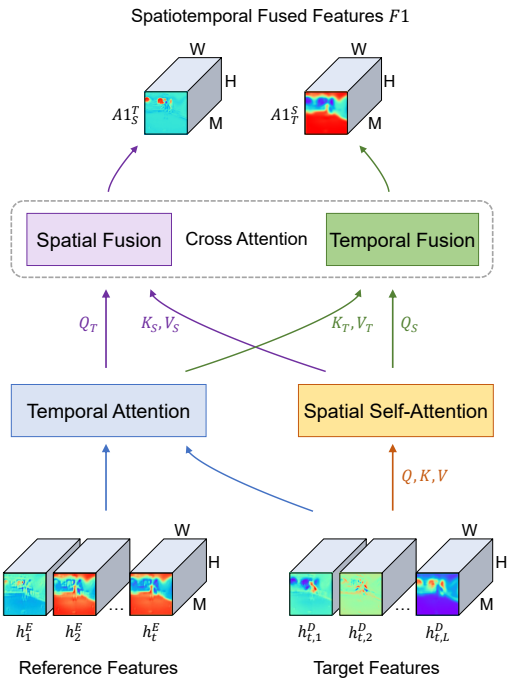


Figure 3. SA module of SRVP. This module performs temporal attention, spatial self-attention, and cross-attention.

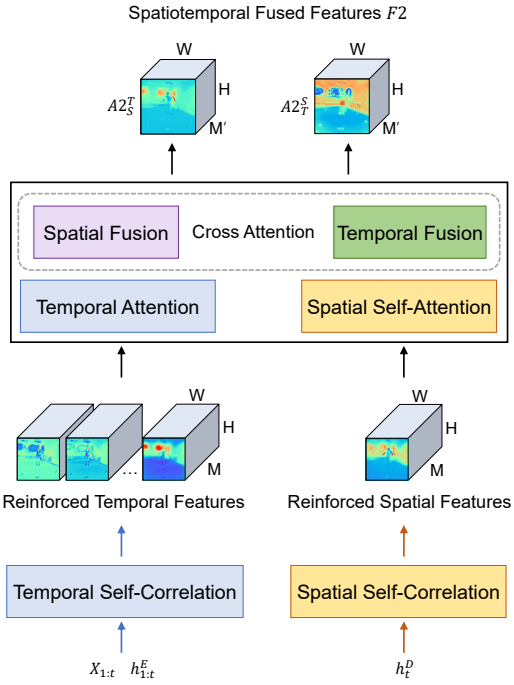


Figure 4. RFA module of SRVP. This module generates reinforced spatiotemporal features and then performs temporal attention, spatial self-attention, and cross-attention.

tions between frames and spatial information within frames. Instead of directly utilizing these feature representations, applying contrast enhancement prior to the attention mechanism enables a more pronounced extraction of features related to object shape. Therefore, we designed the RFA module (Fig. 4), which performs feature reinforcement and subsequently generates the spatiotemporal fused features based on the attention mechanism used in the SA module.

The main methods for constructing SRVP are as follows:

Temporal Attention Our approach employs the conventional scaled dot-product attention mechanism [23]. The target features, $h^D_{t,1:L} \in \mathbb{R}^{L\times MHW}$, are used as the query, whereas the key and value are derived from the reference features, $h^E_{1:t} \in \mathbb{R}^{N\times MHW}$. The temporal attention process is then formulated as follows:

$$\begin{aligned} \omega &= h^D_{t,1:L} \otimes (h^E_{1:t})^\top \\ \mathrm{A}^T &= \mathrm{Softmax}(\frac{\mathrm{Norm}(\omega)}{d}) \otimes h^E_{1:t} \end{aligned} \tag{2}$$

where $\otimes$ denotes matrix multiplication, d is $\sqrt{MHW}$, and $\omega \in \mathbb{R}^{L\times N}$ represents the similarity score between the target and the reference features. The L-wise L2 normalization is applied to ω before applying the softmax function. Through this process, temporal context features $\mathrm{A}^T \in \mathbb{R}^{M\times HW}$ containing the reference feature information most relevant to the currently predicted image are obtained.

Spatial Self-Attention The target features $h^D_{t,1:L}$ are normalized along the L-axis and transformed to the shape of $h^D_{t,n} \in \mathbb{R}^{M\times HW}$. Thereafter, they are embedded into Q, K, and $V \in \mathbb{R}^{M\times HW}$ as follows:

$$\begin{aligned} Q &= W_q(h^D_{t,n}) \\ K &= W_k(h^D_{t,n}) \\ V &= W_v(h^D_{t,n}) \end{aligned} \tag{3}$$

where W_q, W_k, and W_v represent channel-wise (M-wise) linear projections. Similar to temporal attention, we apply scaled dot-product attention to derive the spatial context features $\mathrm{A}^S \in \mathbb{R}^{M\times HW}$:

$$\mathrm{A}^S = \mathrm{Softmax}(\frac{Q \otimes K^\top}{d_k}) \otimes V \tag{4}$$

where d_k is $\sqrt{HW}$. A^S encodes information regarding the regions within the image that should be emphasized, thereby enhancing spatial feature extraction.

Cross-Attention Through the aforementioned processes, A^T captures temporal context information, whereas A^S retains information on short-term variations. Subsequently, these two distinct feature representations are fused to generate dynamic spatiotemporal features. First, a linear projection is applied to A^T and A^S, similar to that in Eq. (3), to

generate the query, key, and value, which are used to compute the spatiotemporally fused features $F1 \in \mathbb{R}^{2M \times HW}$ as follows:

$$\begin{aligned} \mathrm{A1}_S^T &= \mathrm{Softmax}(\frac{Q_T \otimes {K_S}^\top}{d_k}) \otimes V_S \\ \mathrm{A1}_T^S &= \mathrm{Softmax}(\frac{Q_S \otimes {K_T}^\top}{d_k}) \otimes V_T \\ F1 &= [\mathrm{A1}_S^T, \mathrm{A1}_T^S] \end{aligned} \tag{5}$$

where Q_T, K_T, and $V_T \in \mathbb{R}^{M \times HW}$ are produced by linear projections of A^T, whereas Q_S, K_S, and $V_S \in \mathbb{R}^{M \times HW}$ orignate from A^S. $\mathrm{A1}_S^T \in \mathbb{R}^{M \times HW}$ represents a spatially fused temporal context feature and $\mathrm{A1}_T^S \in \mathbb{R}^{M \times HW}$ denotes a temporally fused spatial context feature.

Feature Reinforcement Reinforced features ${h'}_{1:t}^E \in \mathbb{R}^{N \times MHW}$ and ${h'}_t^D \in \mathbb{R}^{M \times HW}$ were generated based on different self-correlation maps [6, 17, 24]. Following self-correlation calculations, the attention process employed in the SA module is used to obtain the spatiotemporally fused features $F2 = [\mathrm{A2}_S^T, \mathrm{A2}_T^S] \in \mathbb{R}^{2M' \times HW}$, as shown in Fig. 4.

Temporal Self-Correlation Reinforced temporal features are generated using both the spatial feature map of historical frames and the reference features. First, convolution is applied to each frame of the historical sequence, producing the feature map $X' \in \mathbb{R}^{N \times M \times H \times W}$ as follows:

$$X' = \{\mathrm{ConvNet}(X_i)\}_{i=1}^t \tag{6}$$

ConvNet includes 2D convolution layers, batch normalization, and ReLU. This process extracts spatial features from each frame. Subsequently, time-wise (N-wise) softmax and L2 normalization are employed to identify the most active feature map within the past frames. Based on this, the temporal self-correlation map $\psi^T \in \mathbb{R}^{N \times MHW}$ is generated as follows:

$$\psi^T = \mathrm{Norm}(X') \otimes \mathrm{Norm}(X' \otimes \mathrm{Softmax}(X')^\top)^\top \tag{7}$$

Finally, the reinforced temporal features ${h'}_{1:t}^E$ are obtained as

$${h'}_{1:t}^E = \mathrm{LayerNorm}(h_{1:t}^E + \psi^T) \tag{8}$$

This process enhances the local contrast of the reference features, enabling the extraction of richer video dynamics. ${h'}_{1:t}^E$ is generated during the initial time-step prediction and remains constant for subsequent predictions.

Spatial Self-Correlation This process focuses on enhancing the spatial details of $h_{t,1:L}^D$. To effectively utilize both low and high-level features, the hidden states across all layers are concatenated into a single dimension, transforming them into the shape of $h_t^D \in \mathbb{R}^{LM \times HW}$. The spatial self-correlation map $\psi^S \in \mathbb{R}^{M \times HW}$ is computed in a similar manner as in Eq. (7):

$$\psi^S = \mathrm{Norm}(h_t^D) \otimes \mathrm{Norm}(h_t^D \otimes \mathrm{Softmax}(h_t^D)^\top)^\top \tag{9}$$

Channel-wise (LM-wise) softmax and L2 normalization are applied in this case. The reinforced spatial features ${h'}_t^D$ are then generated as follows:

$${h'}_t^D = \mathrm{ConvNet}(\mathrm{LayerNorm}(h_t^D + \psi^S)) \tag{10}$$

ConvNet comprises a 2D convolution layer and Tanh activation function, allowing for the extraction of a spatially sharper feature map. Notably, this approach enhances contrast in regions where motion occurs (e.g., a person's limbs).

Output Layer The SA module generates spatiotemporally fused features $F1$ using $h_{1:t}^E$ and $h_{t,1:L}^D$, whereas the RFA module utilizes ${h'}_{1:t}^E$, and ${h'}_t^D$ to compute $F2$. Finally, the output layer predicts X_{t+1} as follows:

$$X'_{t+1} = \mathrm{ConvNet}(\mathrm{Concat}(h_t^D, F1, F2)) \tag{11}$$

ConvNet comprises a 2D convolution layer and sigmoid activation function. In the conventional baseline model (Fig. 1), predictions are generated using only $h_{t,1:L}^D$. In contrast, SRVP uses enhanced spatiotemporal features, such as $F1$ and $F2$, producing more refined predictions.

4. Experiments

4.1. Benchmarks

We assessed the performance of our proposed model using three established benchmark datasets: Moving MNIST [20], KTH Action [18], and Human3.6M [12].

Moving MNIST [20] is a widely used benchmark for VP, comprises 64×64 grayscale image sequences. We followed the procedure in [16] to generate 10,000 training sequences and used the 10,000 test sequences provided in [20]. The models were trained to predict the 10 subsequent frames after receiving 10 input frames. Additionally, we evaluated the models' ability to capture long-term dependencies by extending the prediction length to 30 frames.

KTH Action [18] contains grayscale videos showing various human actions, categorized into six groups. Each category includes 25 subjects, performing the actions under four different conditions: outdoors, outdoors with scale variations, outdoors with different attire, and indoors. Consistent with previous research, we used subjects 1– 16 for training and 17–25 for testing. Images were center-cropped and resized to 64×64. All models were trained to predict the next 10 frames by taking 10 observed frames.

Human3.6M [12] contains color video sequences depicting 17 different human actions. Subjects S1, S5, S6, S7,

and S8 were used for training, whereas S9 and S11 were reserved for testing. Our experiments focused on the Walking and WalkTogether scenarios, with images resized from 1000×1000 to 100×100. The models were trained to predict the next 4 frames by taking 4 observed frames.

4.2. Reference Models

We built five RNN-based VP models by equipping the recurrent layers of the baseline architecture with several previously proposed spatiotemporal memory units, including ConvLSTM [19], ST-LSTM [25], Causal LSTM [26], MIM [28], and ConvGRU (Sec. 3.2). For the Causal LSTM configuration, we followed the method outlined in the original study [26], incorporating a GHU into the first recurrent layer for five layers. To evaluate the ability of SRVP to mitigate the error accumulation problem inherent in RNN-based models, we compared its performance with that of RNN-free models. We selected MIMO-VP [16], which employs a Transformer-based prediction approach, and SimVP [8], which employs a CNN-based method, as our reference models. To the best of our knowledge, these are the most recently published state-of-the-art models.

4.3. Implementation Details

We trained the SRVP and the RNN-based VP models using the binary cross-entropy loss function. The learning rate was selected from the set $\{1e-4, 1e-5, 5e-6\}$, depending on the specific dataset. Model optimization was performed using the RMSProp optimizer and a cosine annealing scheduler. All models were trained in more than 150 epochs, with a batch size of 8. For the RNN-free models, we adhered to the configurations described in their original study. Model performance was evaluated using mean squared error (MSE), peak signal-to-noise ratio (PSNR), and the structural similarity index measure (SSIM). Our implementation is available at https://github.com/yuseonk/SRVP.

4.4. Evaluations

In this section, we conduct both quantitative and qualitative evaluations. All qualitative results are illustrated in the supplementary material.

Moving MNIST This dataset comprised sequences featuring two simultaneously moving digits, identical or different. The primary challenge involved accurately predicting the shapes of the digits, particularly when they overlapped or moved apart. The quantitative results for a 10-frame prediction horizon are presented in Tab. 1. SRVP demonstrated improvements of up to 39.02% in MSE and 36.36% in SSIM compared to traditional RNN-based models. However, its performance was lower than that of the RNN-free models.

Nonetheless, as illustrated in Fig. 5, SRVP significantly reduced the blurring artifacts observed in the traditional RNN-based models, achieving visual quality comparable to the RNN-free models. By incorporating two attention modules, SA and RFA, into ConvGRU, SRVP introduces a modest increase in computational cost (approximately 6.3% in FLOPs) compared to ConvGRU. However, this trade-off is justified by the significant improvement in prediction accuracy (approximately 9.6% in MSE) and enhanced capture of fine-grained motion details. These findings indicate that SRVP more effectively preserves the spatial representation of objects, thereby mitigating the error accumulation problem that affects long-term predictions in RNN-based models.

Furthermore, in the long-term dependency evaluation,

Table 1. Performance comparison on the Moving MNIST dataset. Lower MSE and higher SSIM values indicate better results. The results are averaged for all predicted frames.

Method	10 → 10		10 → 30	
	MSE ↓	SSIM ↑	MSE ↓	SSIM ↑
ConvLSTM [19]	1419.41	0.5661	1272.60	0.4963
ST-LSTM [25]	1533.24	0.5481	1324.94	0.4636
Causal LSTM [26]	1162.65	0.6693	1303.48	0.4886
MIM [28]	1385.35	0.5950	1303.11	0.4908
ConvGRU (Sec. 3.2)	1033.93	0.7082	1257.30	0.5221
MIMO-VP [16]	**401.59**	**0.9049**	**1021.55**	**0.6564**
SimVP [8]	673.01	0.8283	1280.05	0.5167
SRVP	935.02	0.7474	1173.18	0.5533

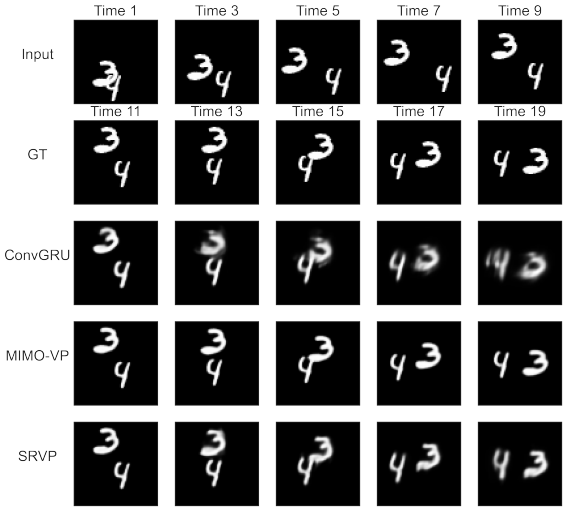


Figure 5. Prediction results for the Moving MNIST (10 → 10). GT indicates ground-truth sequences. ConvGRU is the best-performing RNN-based model, and MIMO-VP is the best-performing RNN-free model.

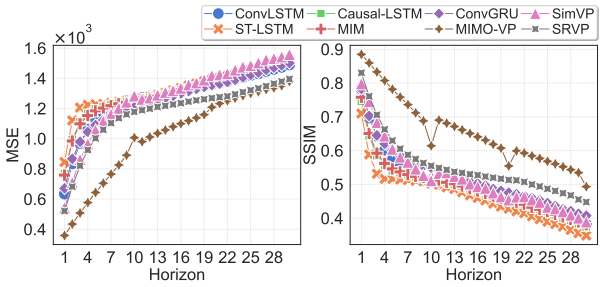


Figure 6. Frame-wise quantitative results for the Moving MNIST dataset (10 → 30).

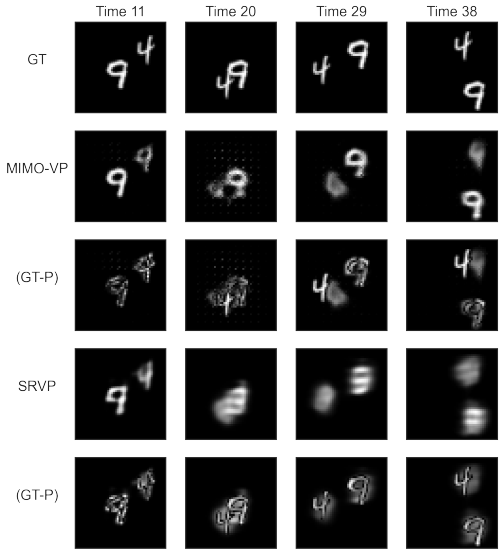


Figure 7. Prediction results for the Moving MNIST dataset (10 → 30). GT represents ground-truth sequences, whereas GT-P indicates the difference between the ground-truth and predicted frames. MIMO-VP shows dot artifacts at Time 20.

SRVP showed improvements of 8.35% in MSE and 7.08% in SSIM compared to SimVP. Although MIMO-VP achieved the best quantitative performance, it made inconsistent predictions over time, with large errors manifesting in the middle of sequences (Figs. 6 and 7). Therefore, SRVP might be more effective than RNN-free models in capturing long-term temporal relationships.

KTH Action The key difficulty with this dataset is estimating human motion over long durations. This requires the model to simultaneously predict changes in position within the scene and the movements of limbs. The quantitative results listed in Tab. 2 reveal that SRVP achieved superior performance on this challenging task. SRVP demonstrated improvements of up to 59.09% in MSE, 29.69% in PSNR, and 38.29% in SSIM compared to reference models.

Fig. 8 provides a qualitative comparison between RNN-free models and SRVP. SRVP effectively captures the overall human motion trend, whereas the RNN-free models fail to accurately predict the trajectory. This leads to considerable errors in the position where movement occurs when performing a pixel-by-pixel comparison. Moreover, efforts are still needed to preserve the spatial details. Future improvements should distinguish the high-motion areas from the background.

Table 2. Performance comparison on the KTH dataset. Lower MSE and higher PSNR and SSIM values indicate better results. The results are averaged for all predicted frames.

Method	MSE ↓	PSNR ↑	SSIM ↑
ConvLSTM [19]	936.69	17.1198	0.5322
ST-LSTM [25]	386.21	21.9665	0.7294
Causal LSTM [26]	408.14	21.5750	0.7158
MIM [28]	445.27	21.3214	0.6947
ConvGRU (Sec. 3.2)	418.09	21.5636	0.7115
MIMO-VP [16]	502.15	21.0167	0.7276
SimVP [8]	452.65	21.4365	0.7179
SRVP	**383.16**	**22.0322**	**0.7360**

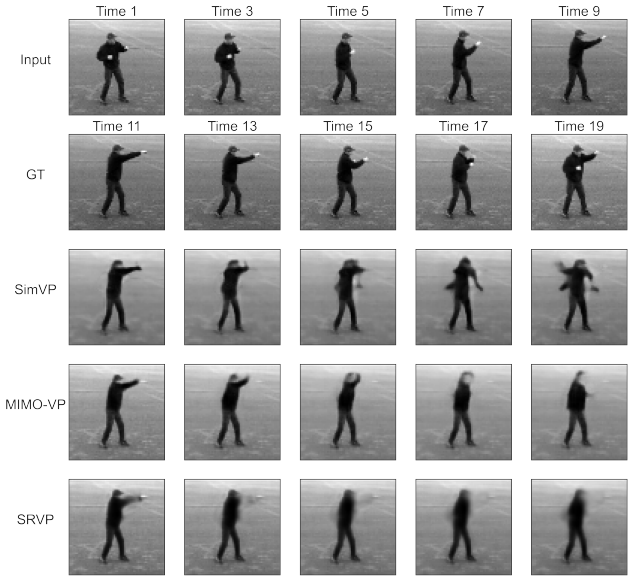


Figure 8. Prediction results for the KTH dataset (10 → 10). GT indicates ground-truth sequences.

Human3.6M The three-dimensional nature of the images in this dataset makes it more challenging than the previous two. Quantitative performance was evaluated for each predicted frame (Tab. 3). ST-LSTM and SRVP exhibit the highest performance, with a minimal difference of approximately 0.1% between them. Furthermore, SRVP outperformed RNN-free models by approximately 1.64 to 3.9%. Although the RNN-free models excelled on less complex datasets such as Moving MNIST, their performance de-

Table 3. Performance comparison for each frame on the Human3.6M dataset. Higher SSIM values indicate better results.

Method	Horizon			
	1	2	3	4
ConvLSTM [19]	0.8960	0.8799	0.8617	0.8440
ST-LSTM [25]	0.9656	**0.9519**	**0.9397**	**0.9285**
Causal LSTM [26]	0.9639	0.9502	0.9375	0.9262
MIM [28]	0.9599	0.9446	0.9313	0.9185
ConvGRU (Sec. 3.2)	0.9643	0.9489	0.9354	0.9235
MIMO-VP [16]	0.9303	0.9154	0.9034	0.8924
SimVP [8]	0.9415	0.9286	0.9194	0.9122
SRVP	**0.9659**	0.9512	0.9386	0.9272

Table 4. Ablation results for the Moving MNIST dataset (10 $\rightarrow$ 10). Without-CrossAtt indicates the results obtained after removing the cross-attention process in both SA and RFA.

Method	MSE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$
SRVP	935.02	18.6160	0.7474
Without-SA	968.64	18.4719	0.7365
Without-RFA	**1023.84**	**18.2027**	**0.6983**
Without-CrossAtt	973.16	18.4690	0.7314

clined as the image size and spatial complexity increased.

Fig. 9 presents a qualitative comparison between the best-performing RNN-based model, the best-performing RNN-free model, and SRVP. The frames predicted by SRVP more accurately retain the human shape than those predicted by ST-LSTM, with reduced residual artifacts on the moving arms. In contrast, SimVP exhibits inconsistent image texture and significant errors, not only in the human figure but also in the background.

Ablation Study We also conducted an ablation study to assess the individual contribution of the proposed attention modules. Tab. 4 presents the results, indicating that the RFA module is the most critical component of SRVP. This finding suggests that performing attention based on spatiotemporally enhanced contrast features is more effective than directly using hidden state information from recurrent memory.

5. Conclusions

This paper presents SRVP, a VP model designed to enhance object representation retention for improved long-term forecasting. SRVP integrates two attention-based modules, SA and RFA, to jointly capture temporal dependencies and spatial correlations. This dual-attention mechanism mitigates the common issue of object detail degradation in long-term

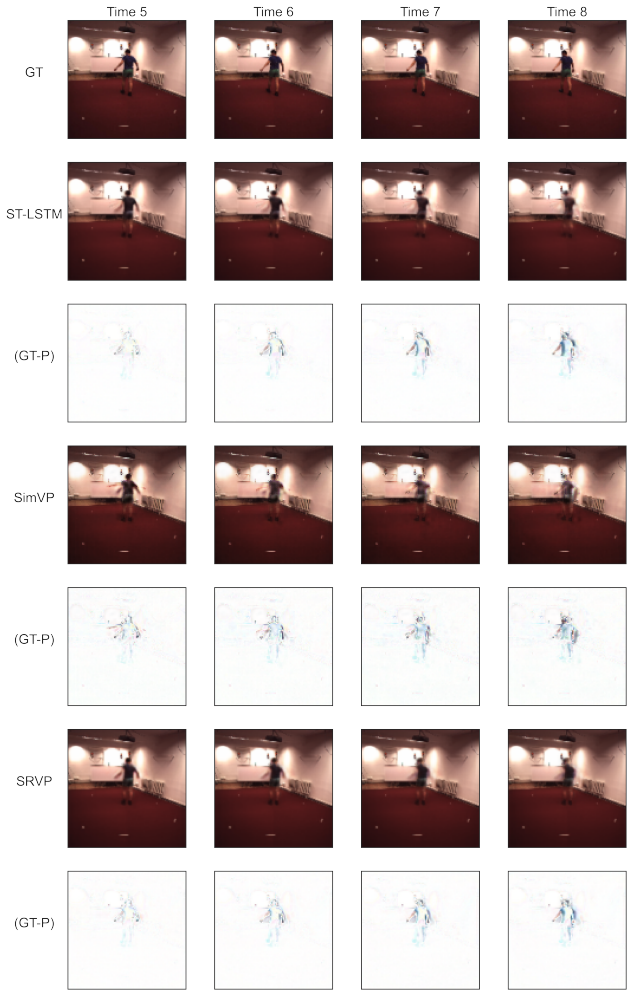


Figure 9. Prediction results for the Human3.6M dataset (4 $\rightarrow$ 4). GT represents the ground-truth sequences, whereas GT-P denotes the difference between the ground-truth and predicted frames.

VP, where moving objects often blur over time. Empirical evaluations on three benchmark datasets demonstrate that SRVP effectively reduces image degradation in RNN-based models while achieving performance comparable to RNN-free architectures. These results underscore the benefits of our attention mechanisms in improving VP model accuracy and robustness. Future research should explore integrating segmentation techniques to refine pixel-level object representations further.

Acknowledgments

This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP, under grant number RS-2024-00397359) and the Korea Institute of Science and Technology Information (KISTI, under grant number K25L1M2C2-01).

References

[1] Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, pages 1–15, 2015. 3

[2] Adrian Bulat, Juan Manuel Perez Rua, Swathikiran Sudhakaran, Brais Martinez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. In *NeurIPS*, pages 19594–19607, 2021. 1

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 1

[4] Zheng Chang, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. Stam: A spatiotemporal attention based memory for video prediction. *IEEE TMM*, 25:2354–2367, 2022. 1

[5] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014. 2, 3

[6] Suhwan Cho, Minhyeok Lee, Seunghoon Lee, Dogyoon Lee, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Dual prototype attention for unsupervised video object segmentation. In *CVPR*, pages 19238–19247, 2024. 5

[7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014. 3

[8] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler yet better video prediction. In *CVPR*, pages 3170–3180, 2022. 6, 7, 8

[9] Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10):2451–2471, 2000. 1

[10] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *CVPR*, pages 11474–11484, 2020. 1

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 1

[12] 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 TPAMI*, 36(7):1325–1339, 2014. 5

[13] Zhihui Lin, Maomao Li, Zhuobin Zheng, Yangyang Cheng, and Chun Yuan. Self-attention convlstm for spatiotemporal prediction. In *AAAI*, pages 11531–11538, 2020. 1

[14] Zhouyong Liu, Shun Luo, Wubin Li, Jingben Lu, Yufan Wu, Shilei Sun, Chunguo Li, and Luxi Yang. Convtransformer: A convolutional transformer network for video frame synthesis. *arXiv preprint arXiv:2011.10185*, 2020. 1

[15] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

[16] Shuliang Ning, Mengcheng Lan, Yanran Li, Chaofeng Chen, Qian Chen, Xunlai Chen, Xiaoguang Han, and Shuguang Cui. Mimo is all you need: A strong multi-in-multi-out baseline for video prediction. In *AAAI*, pages 1975–1983, 2023. 1, 5, 6, 7, 8

[17] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. Unveiling the potential of structure preserving for weakly supervised object localization. In *CVPR*, pages 11642–11651, 2021. 5

[18] Christian Schuldt., Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004. 5

[19] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *NeurIPS*, 28, 2015. 1, 2, 6, 7, 8

[20] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *ICML*, pages 843–852. PMLR, 2015. 1, 5

[21] Jiahao Su, Wonmin Byeon, Jean Kossaifi, Furong Huang, Jan Kautz, and Anima Anandkumar. Convolutional tensor-train lstm for spatio-temporal learning. *NeurIPS*, 33:13714–13726, 2020. 1, 2

[22] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NeurIPS*, page 3104–3112, 2014. 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, page 6000–6010, 2017. 3, 4

[24] Linhan Wang, Shuo Lei, Jianfeng He, Shengkun Wang, Min Zhang, and Chang-Tien Lu. Self-correlation and cross-correlation learning for few-shot remote sensing image semantic segmentation. In *Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, pages 1–10, 2023. 5

[25] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *NeurIPS*, 30, 2017. 1, 2, 6, 7, 8

[26] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *ICML*, pages 5123–5132. PMLR, 2018. 1, 2, 6, 7, 8

[27] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3d lstm: A model for video prediction and beyond. In *ICLR*, 2018.

[28] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *CVPR*, pages 9154–9162, 2019. 1, 2, 6, 7, 8

[29] Haixu Wu, Zhiyu Yao, Jianmin Wang, and Mingsheng Long. Motionrnn: A flexible model for video prediction with spacetime-varying motions. In *CVPR*, pages 15435–15444, 2021. 1

[30] Xi Ye and Guillaume-Alexandre Bilodeau. Video prediction by efficient transformers. *Image and Vision Computing*, 130: 104612, 2023. 1

[31] Yuhui Yuan, Rao Fu, Lang Huang, Weihong Lin, Chao Zhang, Xilin Chen, and Jingdong Wang. Hrformer: High-resolution transformer for dense prediction. In *NeurIPS*, pages 7281–7293, 2021. 1