VIDEO ANOMALY DETECTION VIA PREDICTION NETWORK WITH ENHANCED SPATIO-TEMPORAL MEMORY EXCHANGE

Guodong Shen, Yuqi Ouyang, and Victor Sanchez

Department of Computer Science, University of Warwick, Coventry, UK

ABSTRACT

Video anomaly detection is a challenging task because most anomalies are scarce and non-deterministic. Many approaches investigate the reconstruction difference between normal and abnormal patterns, but neglect that anomalies do not necessarily correspond to large reconstruction errors. To address this issue, we design a Convolutional LSTM Auto-Encoder prediction framework with enhanced spatio-temporal memory exchange using bi-directionalilty and a higher-order mechanism. The bi-directional structure promotes learning the temporal regularity through forward and backward predictions. The unique higher-order mechanism further strengthens spatial information interaction between the encoder and the decoder. Considering the limited receptive fields in Convolutional LSTMs, we also introduce an attention module to highlight informative features for prediction. Anomalies are eventually identified by comparing the frames with their corresponding predictions. Evaluations on three popular benchmarks show that our framework outperforms most existing prediction-based anomaly detection methods.

Index Terms— video anomaly detection, prediction, ConvLSTM Auto-Encoder, bi-directional, attention

1. INTRODUCTION

Video anomaly detection aims to detect abnormal patterns in motion and appearance. The main challenges of this task come from the rarity and diversity of anomalies in the real world. To this end, recent works adopt unsupervised learning algorithms, i.e., modeling normal patterns from a training set and then discriminating anomalies based on the discrepancy from these normal patterns. One representative strategy is measuring the reconstruction error produced by Auto-Encoders (AEs) under the assumption that anomalies would lead to outliers in the image or sparse coding space [1, 2]. However, despite its prevalence, this strategy exposes two significant limitations. First, the assumption may not necessarily hold due to the high capacity of neural networks to reconstruct anomalies precisely. Second, the temporal dynamics of videos, which help to uncover motion irregularities, are usually confined to alternative representations, e.g., optical flow.

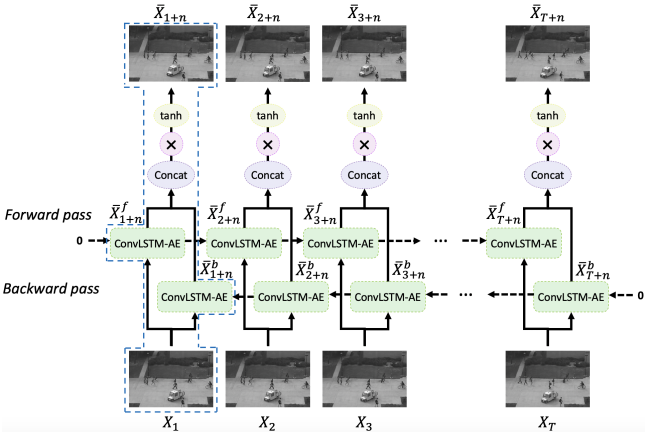

Fig. 1: The pipeline of the proposed framework (unfolded over time). $\otimes$ denotes the convolution operation. The dashed line highlights its fundamental structure.

Prediction-based approaches can overcome the aforementioned limitations by considering anomalies as unexpected events. Specifically, during training, they enforce a model to generate accurate predictions (i.e., future frames) for normal events. Once trained, they infer anomalies from the dissimilarity between prediction and ground truth. This process exploits both underlying spatial and temporal dynamics from historical observations so that anomalies result in poor predictions.

In general, prediction-based approaches can be categorized into two classes. 1) U-Net [3] based methods, which commonly stack consecutive frames to form spatio-temporal cubes that serve as input to U-Nets to predict subsequent frames [4–6]. Naively stacking frames, however, can weaken the correlation between adjacent frames and burden the encoders in U-Nets. 2) Convolutional Long Short-term Memory (ConvLSTM) [7] based methods. Inspired by natural language processing (NLP), these methods employ recurrent frameworks to sequentially process data to generate future frames [8–11]. However, most of these methods arrange and stack ConvLSTM layers like traditional LSTM layers, thus neglecting their difference in receptive fields. Moreover, both classes of methods usually require either adversarial learning [4, 11] or extra reconstruction tasks [5, 8] to attain strong performance.

Within the context of ConvLSTM-based methods, this paper further bridges the gap between NLP and prediction-based video anomaly detection by presenting a spatio-temporal memory-enhanced ConvLSTM-AE framework using bi-directionalilty and a higher-order mechanism [12]. Compared to the uni-directionality adopted by previous works, the bi-directionality supports the reverse processing of the data. Our framework is thus more responsive to the temporal dimension of videos, which typically reflects motion irregularities. Motivated by the higher-order mechanism and PredRNN in [13], we devise a novel spatial higher-order ConvLSTM to boost spatial information exchange between the encoder and the decoder. Furthermore, to cope with the inflexible receptive fields in ConvLSTMs, we also introduce an effective attention module to dynamically highlight the features that are more informative for predicting future frames. Compared to the state-of-the-art (SOTA), the proposed framework demonstrates superior performance through experiments on anomaly detection benchmarks and ablation studies.

2. PROPOSED FRAMEWORK

The pipeline of our framework is illustrated in Fig. 1. It leverages a Bi-LSTM framework [14] as the backbone, where two enhanced ConvLSTM-AEs are used to perform frame-to-frame prediction in forward and backward order. The outputs of both ConvLSTM-AEs are then merged to yield a more precise prediction. The anomalies are eventually identified based on prediction errors.

As shown in Fig. 2, each ConvLSTM-AE consists primarily of Conv, Deconv and ConvLSTM layers that discover temporal dependencies on spatial data. They are stacked alternately to capture long-range features. Our spatial higher-order ConvLSTMs, hereinafter referred to as SHO-ConvLSTM, replace the original ConvLSTMs in the decoder to incorporate hidden states from the current encoder. In addition, the attention modules are implemented before ConvLSTM layers to accentuate the dynamics between frames and overcome the drawbacks of the restricted receptive fields in ConvLSTMs.

Bi-directionality. Using a rigid temporal order in recurrent frameworks can potentially reduce the flexibility of temporal memory interaction [15]. To lift the restriction on temporal order, we introduce a bi-directional mechanism derived from the original Bi-LSTM [14] in NLP. While the traditional Bi-LSTMs rely on past and future data, our mechanism focuses more on the bi-directionality of past frames for real-time use. As shown in Fig. 1, the input sequence $\{X_1, X_2, ..., X_T\}$ of length T flows into two ConvLSTM-AEs, which conduct the forward and backward passes separately. The forward pass propagates the hidden states from timestep 1 to T and sequentially generates the forward predictions $\{\bar{X}^f_{1+n}, \bar{X}^f_{2+n}, ..., \bar{X}^f_{T+n}\}$, where n denotes the n-th frame after the input frame. Conversely, the backward pass conveys the hidden states from timestep T to 1 and generates the predictions in reverse order, $\{\bar{X}^b_{T+n}, \bar{X}^b_{T+n-1}, ..., \bar{X}^b_{1+n}\}$. Both predictions are then concatenated along the channel dimension and mapped back to the image space to produce the final predictions $\{\bar{X}_{1+n}, \bar{X}_{2+n}, ..., \bar{X}_{T+n}\}$. One can consider the process as implicit multi-task learning, where the framework is trained to capture the temporal dimension in two complementary tasks, concurrently.

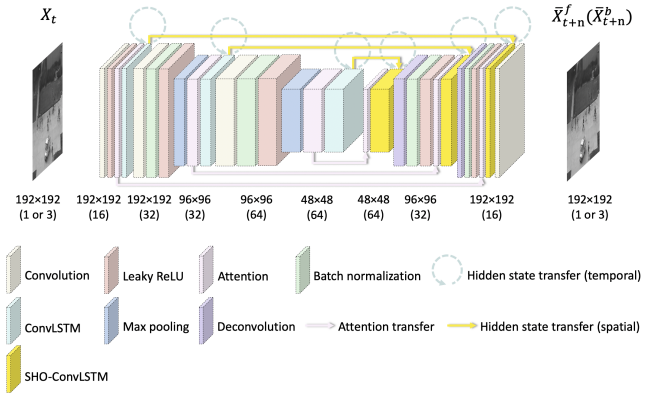

Fig. 2: The architecture of the proposed ConvLSTM-AE. X_t is a frame at timestep t. The output of the ConvLSTM-AE can be either $\bar{X}^f_{t+n}$ in forward pass or $\bar{X}^b_{t+n}$ in backward pass.

SHO-ConvLSTM. Due to the first-order Markovian nature, the traditional ConvLSTM only allows hidden states to be transferred over time but spatially independently, ignoring that spatial transitions can be equivalently essential to generate future frames [13]. Thus, we derive the SHO-ConvLSTM to enable spatial memory flows (see Fig. 3), which is formulated as follows:

$$
\begin{aligned}
\bar{C}^d_t &= \tanh\left(W_c \otimes \left[\tilde{X}^d_t, H^d_{t-1}, \boldsymbol{H^e_t}\right] + b_c\right), \\
i^d_t &= \sigma\left(W_i \otimes \left[\tilde{X}^d_t, H^d_{t-1}, \boldsymbol{H^e_t}\right] + b_i\right), \\
f^d_t &= \sigma\left(W_f \otimes \left[\tilde{X}^d_t, H^d_{t-1}, \boldsymbol{H^e_t}\right] + b_f\right), \\
o^d_t &= \sigma\left(W_o \otimes \left[\tilde{X}^d_t, H^d_{t-1}, \boldsymbol{H^e_t}\right] + b_o\right), \\
C^d_t &= f^d_t \circ C^d_{t-1} + i^d_t \circ \bar{C}^d_t, \\
H^d_t &= o^d_t \circ \tanh\left(C^d_t\right),
\end{aligned}
\tag{1}
$$

where $\tilde{X}^d_t$ is the input to the SHO-ConvLSTM at timestep t; H^d_{t-1} is the hidden state from timestep $t-1$; H^e_t is the hidden state from the corresponding ConvLSTM (of the same size) in the encoder at timestep t; $\bar{C}^d_t$, i^d_t, f^d_t and o^d_t denote the cell, input, forget and output states at timestep t, respectively; $\sigma(\cdot)$ denotes the Sigmoid function; $\circ$ and $\otimes$ denote the Hadamard product and convolution operations, respectively; and W and b are trainable parameters of the SHO-ConvLSTM. The symbols in bold highlight the difference between a ConvLSTM and the SHO-ConvLSTM.

As shown in Fig. 3, while the ConvLSTM in the encoder still follows the standard ConvLSTM routine, the SHO-

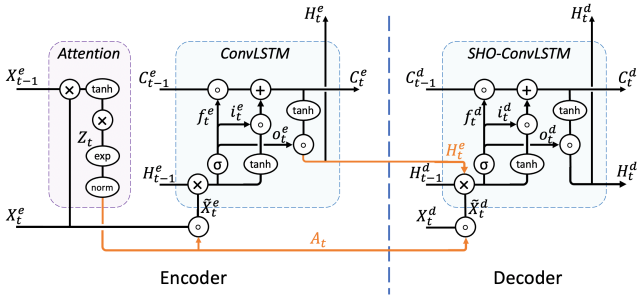

Fig. 3: The spatial higher-order mechanism and attention module. Superscripts ^{e} and ^{d} denote elements of the encoder and decoder, respectively; H_t^e denotes the spatial memory flow between ConvLSTM and SHO-ConvLSTM; A_t denotes the attention mask.

ConvLSTM in the decoder considers both the temporal dynamics H_{t-1}^d and the spatial dynamics H_t^e. Hence, the decoder can efficiently retrieve information from both historical timesteps and the current encoder, as well as alleviate the gradient vanishing problem.

Attention. Since ConvLSTMs alone cannot fulfill informative feature selection [16], we introduce an auxiliary attention module to estimate the importance weights of different elements in the feature maps (see Fig. 3). The intuition behind the module is that the anomalies of interest typically cause distinctive variations in pixel values over time. To focus on the variations, we compute the attention weights Z_t by encoding both the current input X_t^e and the previous input X_{t-1}^e:

$$Z_t = W_z^2 \otimes \tanh\left(W_z^1 \otimes [X_t^e, X_{t-1}^e] + b_z\right). \tag{2}$$

Different from the schemes of convolving X_t^e with the hidden state H_{t-1}^e [16,17], the proposed module does not rely on the ConvLSTM's output, and thus can achieve parallel computing. Meanwhile, it eases feature fusion and extraction by taking the inputs with the identical context definition. The final attention mask A_t is then generated for each feature map using *min-max* normalization:

$$A_t^{ij} = \frac{\exp\left(Z_t^{ij}\right) - \min_{i,j}\left(\exp\left(Z_t^{ij}\right)\right)}{\max_{i,j}\left(\exp\left(Z_t^{ij}\right)\right) - \min_{i,j}\left(\exp\left(Z_t^{ij}\right)\right)}, \tag{3}$$

where (i, j) represents the element's position in the map; the $\max_{i,j}(\cdot)$ and $\min_{i,j}(\cdot)$ operations, respectively, locate the maximum and minimum elements in Z_t. Compared to *sum*, i.e., *Softmax*, or *max-only* normalization [17], *min-max* allows for multiple maxima and minima, and enlarges the disparity between them. We hence expect the features that contribute more to the prediction to gain larger weights and vice versa.

As shown in Fig. 3, the attention mask A_t is computed from the feature maps in the encoder and then imposed on both the ConvLSTM and the SHO-ConvLSTM using the Hadamard product, assuming that regions of interest only shift slightly in time and space.

***SSIM* + ℓ_1 loss.** The classic Manhattan (ℓ_1) loss encourages sparsity in nonstructural features and linearly mitigates the impact of outliers:

$$\mathcal{L}^{\ell_1}(P, \hat{P}) = \frac{1}{|P|} \sum_{i,j \in P} |p_{ij} - \hat{p}_{ij}|, \tag{4}$$

where P and $\hat{P}$ denote the ground-truth and its prediction, respectively; p_{ij} and $\hat{p}_{ij}$ are the pixel values of P and $\hat{P}$ at position (i, j), respectively; and $|P|$ is the number of pixels. However, the ℓ_1 loss minimizes uncertainty by averaging all the probable solutions, i.e., by blurring the predictions. The structural similarity index measure *(SSIM)* loss [18], on the other hand, evaluates perceivable differences in structure, luminance and contrast, and thus promotes visually convincing predictions:

$$\mathcal{L}^{ssim}(P, \hat{P}) = \frac{\left(2\mu_P\mu_{\hat{P}} + c_1\right)\left(2\sigma_{P\hat{P}} + c_2\right)}{\left(\mu_P^2 + \mu_{\hat{P}}^2 + c_1\right)\left(\sigma_P^2 + \sigma_{\hat{P}}^2 + c_2\right)}, \tag{5}$$

where c_1 and c_2 are the stabilizers; μ_P and $\mu_{\hat{P}}$ are the expectations of P and $\hat{P}$, respectively; σ_P^2 and $\sigma_{\hat{P}}^2$ are the corresponding variances; and $\sigma_{P\hat{P}}$ is the covariance. Unfortunately, unlike the ℓ_1 loss, the *SSIM* loss suffers from the insensitivity to nonstructural distortions (e.g., shifting, scaling).

To benefit from both losses, we merge them to obtain the final objective function:

$$\mathcal{L}^{mix} = \mathcal{L}^{ssim} + \lambda(W_{\ell_1} \otimes \mathcal{L}^{\ell_1} + b_{\ell_1}), \tag{6}$$

where the weight λ is pragmatically set to 1. Different from [19], an extra Gaussian filter parameterized by W_{ℓ_1} and b_{ℓ_1} is applied to the ℓ_1 loss to ensure coherence with the Gaussian-based *SSIM* process.

Anomaly inference. To estimate anomaly scores, we adopt the Mean Absolute Error (*MAE*, i.e., ℓ_1) metric as it partially conforms to our objective function. The *MAE* value is calculated for each frame t using Eq. 4 and then normalized over each video to the range $[0, 1]$ by using:

$$S(t) = \frac{MAE(t) - \min_t\left(MAE(t)\right)}{\max_t\left(MAE(t)\right) - \min_t\left(MAE(t)\right)}, \tag{7}$$

where $\max_t(\cdot)$ and $\min_t(\cdot)$ determine the extreme *MAE* values in each video. Large $S(t)$ values correspond to abnormal frames.

3. EXPERIMENTAL RESULTS

Implementation details. We evaluate our framework on three public benchmarks: the UCSD Ped2 [20], the CUHK Avenue [21], and the ShanghaiTech [22] datasets. All input frames are re-sized to 192×192 and normalized to the intensity range $[-1, 1]$. To verify the framework's robustness to different video sources, we preserve the original color format of each dataset: grayscale and RGB. The length T of each sample clip is fixed to 9 since anomalies usually occur in the short term. The receptive fields of the attention, Conv and

Table 1: AUC (%) of different models on three public benchmarks.

Authors	UCSD Ped2	CUHK Avenue	ShanghaiTech
Luo et al. [8]	88.1	77.0	-
Luo et al. [22]	92.2	81.7	68.0
Wang et al. [9]	88.9	**90.3**	-
Liu et al. [4]	95.4	85.1	72.8
Lee et al. [10]	**96.6**	90.0	**76.2**
Song et al. [11]	90.3	89.2	70.0
Chen et al. [6]	**96.6**	87.8	-
Lai et al. [5]	95.8	87.4	-
Our framework	**98.3**	**90.7**	**79.7**

Deconv layers are set to 3×3. The receptive fields of ConvLSTM and SHO-ConvLSTM layers are set to 5×5 [7]. Leaky ReLU is used as the activation function. Batch normalization is inserted right after Conv and Deconv layers as the stabilizer. Adam, with an initial learning rate of 5.0×10^{-4}, serves as the parameter optimizer. We create a validation set from the original training set with a ratio $1:9$, and execute early stopping when the validation loss reaches an optimal value. To differentiate the prediction ability from the reconstruction ability, instead of simply predicting the next frame [4, 5], our framework predicts the 7th frame after each input frame ($n = 7$) and regards the 5th frame in the prediction sequence (i.e., the 3rd frame after the last input frame) as the expected result during testing.

Metrics. We utilize the frame-level Area Under Curve (AUC) of the Receiver Operating Characteristics (ROC) curve to quantitatively assess our framework against the SOTA. A higher AUC value suggests a higher probability that the framework will correctly detect abnormal frames.

Comparison with the SOTA. As shown in Table 1, we compare our framework with several prediction-based SOTA methods. Note that only unsupervised approaches are considered since supervised methods usually acquire knowledge from testing videos, which contradicts typical application scenarios. It is evident that the proposed framework outperforms most competitors on all three benchmarks. Even on the most challenging ShanghaiTech dataset, it reports an overall score of 79.7%, exceeding the SOTA performance by 3.5%.

Ablation studies. We study the impact of different components in our framework: bi-directionality, spatial higher-order mechanism, and attention module. To this end, we create several variants of the proposed framework by removing these components. The experiments are conducted on the CUHK Avenue dataset and assessed in terms of convergence (i.e., optimal validation loss) and AUC. As shown in Table 2, each component contributes individually to the framework's performance for both metrics. Particularly, the bi-directionality and spatial higher-order mechanism significantly enhance the AUC score by 2.6% ($\mathbb{A}$ vs. $\mathbb{D}$) and 2.1% ($\mathbb{B}$ vs. $\mathbb{D}$), respectively, proving the effectiveness of strengthening memory exchange in ConvLSTMs.

Fig. 4 depicts qualitative examples of the attention module's responses to different frames. By analyzing the error maps between prediction and ground truth, we observe that our attention mechanism can diminish the prediction error around normal moving items, while preserving it in abnormal areas. This confirms that the attention module can help the framework to focus on more predictive features and correctly predict normal patterns.

Table 2: Evaluation of different components of our framework on the CUHK Avenue dataset.

Index	Framework design			Val loss ($\times 10^{-3}$)	AUC (%)
	Bi	SHO	Att		
$\mathbb{A}$	✗	✓	✓	17.58	88.1
$\mathbb{B}$	✓	✗	✓	10.49	88.6
$\mathbb{C}$	✓	✓	✗	4.62	89.8
$\mathbb{D}$	✓	✓	✓	4.54	90.7

Bi = Bi-directionality. SHO = Spatial higher-order. Att = Attention. Val loss = Validation loss.

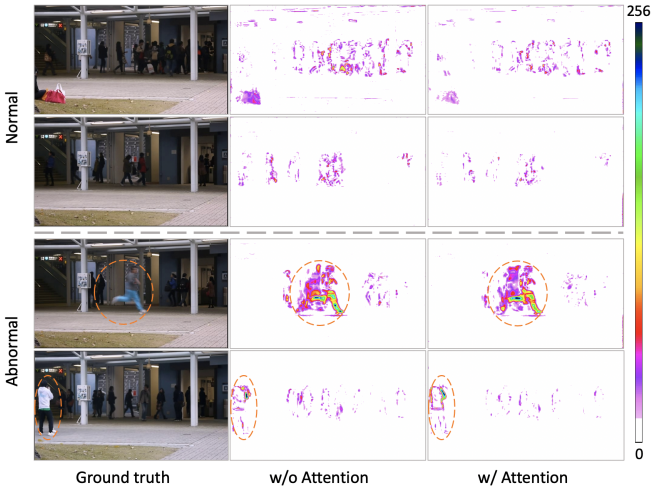

Fig. 4: Examples of the attention module's impact on normal and abnormal frames. The error maps in the second and third column are computed using |prediction − ground truth|. The dashed circles highlight the anomalies.

4. CONCLUSION

We presented a spatio-temporal memory-enhanced ConvLSTM-AE predictor to detect anomalies as the patterns that deviate from expectations. The framework's bi-directionality helps to effectively exploit the temporal dimension through the backward prediction task. Moreover, its spatial higher-order ConvLSTMs can retrieve hidden states from the current encoder to boost spatial information exchange. Finally, its attention module helps to locate the features that are more informative for frame generation by comparing neighboring frames. Our ablation studies and experiments on public benchmarks confirm the effectiveness of each component and validate the strong performance of our framework.

5. REFERENCES

[1] M. Zaheer, J. Lee, M. Astrid, and S. Lee, "Old is gold: Redefining the adversarially learned one-class classifier training paradigm," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14183–14193.

[2] Y. Ouyang and V. Sanchez, "Video anomaly detection by estimating likelihood of representations," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 8984–8991.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[4] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6536–6545.

[5] Y. Lai, R. Liu, and Y. Han, "Video anomaly detection via predictive autoencoder with gradient-based attention," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[6] D. Chen, P. Wang, L. Yue, Y. Zhang, and T. Jia, "Anomaly detection in surveillance video based on bidirectional prediction," *Image and Vision Computing*, vol. 98, pp. 103915, 2020.

[7] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

[8] W. Luo, W. Liu, and S. Gao, "Remembering history with convolutional lstm for anomaly detection," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2017, pp. 439–444.

[9] L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan, "Abnormal event detection in videos using hybrid spatio-temporal autoencoder," in *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 2276–2280.

[10] S. Lee, H. Kim, and Y. Ro, "Bman: bidirectional multi-scale aggregation networks for abnormal event detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 2395–2408, 2019.

[11] H. Song, C. Sun, X. Wu, M. Chen, and Y. Jia, "Learning normal patterns via adversarial attention-based autoencoder for abnormal event detection in videos," *IEEE Transactions on Multimedia*, vol. 22, no. 8, pp. 2138–2148, 2019.

[12] R. Yu, S. Zheng, A. Anandkumar, and Y. Yue, "Long-term forecasting using tensor-train rnns," *Arxiv*, 2017.

[13] Y. Wang, M. Long, J. Wang, Z. Gao, and P. Yu, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 879–888.

[14] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.

[15] M. Schuster and K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[16] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. Snoek, "Videolstm convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.

[17] L. Zhang, G. Zhu, L. Mei, P. Shen, S. Shah, and M. Bennamoun, "Attention in convolutional lstm for gesture recognition," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 2018, pp. 1957–1966.

[18] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[19] H. Zhao, O. Gallo, I. Frosio, and J. Kautz, "Loss functions for image restoration with neural networks," *IEEE Transactions on computational imaging*, vol. 3, no. 1, pp. 47–57, 2016.

[20] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1975–1981.

[21] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2720–2727.

[22] W. Luo, W. Liu, and S. Gao, "A revisit of sparse coding based anomaly detection in stacked rnn framework," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 341–349.