electronics

Article

Vision-Based Prediction of Flashover Using Transformers and Convolutional Long Short-Term Memory Model

M. Hamed Mozaffari *, Yuchuan Li, Niloofar Hooshyaripour and Yoon Ko *

Construction Research Center, National Research Council Canada, 1200 Montreal Road,
Ottawa, ON K1A 0R6, Canada

* Correspondence: mhamed.mozaffarimaaref@nrc-cnrc.gc.ca (M.H.M.); yoon.ko@nrc-cnrc.gc.ca (Y.K.)

Abstract: The prediction of fire growth is crucial for effective firefighting and rescue operations. Recent advancements in vision-based techniques using RGB vision and infrared (IR) thermal imaging data, coupled with artificial intelligence and deep learning techniques, have shown promising solutions to be applied in the detection of fire and the prediction of its behavior. This study introduces the use of Convolutional Long Short-term Memory (ConvLSTM) network models for predicting room fire growth by analyzing spatiotemporal IR thermal imaging data acquired from full-scale room fire tests. Our findings revealed that SwinLSTM, an enhanced version of ConvLSTM combined with transformers (a deep learning architecture based on a new mechanism called multi-head attention) for computer vision purposes, can be used for the prediction of room fire flashover occurrence. Notably, transformer-based ConvLSTM deep learning models, such as SwinLSTM, demonstrate superior prediction capability, which suggests a new vision-based smart solution for future fire growth prediction tasks. The main focus of this work is to perform a feasibility study on the use of a pure vision-based deep learning model for analysis of future video data to anticipate behavior of fire growth in room fire incidents.

Keywords: convolutional long short-term memory; ConvLSTM; deep convolutional neural network; smart firefighting; thermal infrared image analysis; vision-based flashover prediction

Citation: Mozaffari, M.H.; Li, Y.; Hooshyaripour, N.; Ko, Y. Vision-Based Prediction of Flashover Using Transformers and Convolutional Long Short-Term Memory Model. *Electronics* **2024**, *13*, 4776. https://doi.org/10.3390/electronics13234776

Academic Editor: Ping-Feng Pai

Received: 25 October 2024
Revised: 28 November 2024
Accepted: 29 November 2024
Published: 3 December 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

On average, 40,000 fire incidents occurred annually in Canada from 2010 to 2015 [1]. Of these fire incidents, around 19,000 fires were structural fires, including room fires in residential, industrial, and institutional buildings [1]. This means that, in Canada alone, approximately 50 room fires occur every day. Room fires grow fast mainly because of a high amount of petroleum-based fuels (e.g., modern upholstered furniture, carpets, curtains, and plastics in appliances) [2]. In general, room fires are usually more dangerous than fire incidents in open spaces because of the rapid smoke and heat propagation that result in untenable conditions, endangering the occupants in the building. One particularly deadly fire growth phenomenon that usually occurs in room fire incidents is called flashover, when all of the contents in the room ignite simultaneously [3,4]. The ability to detect and predict flashover has been of great importance for fire safety researchers and fire brigades to save the lives of occupants and firefighters. In general, looking at vision-based real-time detection and prediction of flashover is a challenging task [5]. However, significant attempts have been made utilizing various ideas and techniques to detect and predict flashover in-room fire tests. For a list of recent vision-based flashover detection techniques, refer to [2].

Historically, research on fire-growth prediction and detection has relied on data collected from temperature sensors, heat flux gauges, and heat release rate (HRR) calculations [6,7]. However, the practicality of implementing these data collection methods from research labs into actual fire ground is limited, as they require multiple fire-resistant sensors readily installed in specific room locations [8,9]. Advances in computer vision, imaging technology, and artificial intelligence (AI) have led researchers to explore vision-based methods using thermal infrared (IR) and RGB cameras in processing smoke and fire image data as alternatives to point measurement sensors [2,10–14]. AI has played a pivotal role in analyzing the image data of smoke and flame, providing faster analysis and less manual data processing.

From recent advancements in AI, Deep Convolutional Neural Networks (DCNNs) have shown great promise in computer vision tasks [15–17]. Although DCNNs are powerful at finding patterns in spatial data (in applications such as fire and smoke image classification, detection, and segmentation), they fall short in analyzing spatiotemporal data (i.e., videos of fire growth). This limitation has led to the research and development of predictive AI models that aim to forecast future events using time series and sequential data [18]. Techniques like Recurrent Neural Networks (RNNs) have been used to analyze temporal data, but they struggle with long-term dependencies and forget previous information that is crucial for making future predictions [19].

To mitigate this limitation, Long Short-Term Memory (LSTM) networks were introduced. LSTM presented a significant improvement in capturing temporal relationships in data [20]. Among these models, the convolutional LSTM (ConvLSTM) model combines the spatial data analysis strengths of DCNNs with the temporal prediction capabilities of LSTM, offering a robust solution for predicting future events in spatiotemporal higher-dimensional data, such as video data [21]. Nowadays, AI methods leverage the ability of Transformers, a class of deep learning models that have fundamentally changed the landscape of natural language processing (NLP) and have increasingly been applied to various tasks in computer vision, audio processing, and beyond [22]. Transformers use a mechanism called "self-attention" to process input data in parallel, making it highly efficient and scalable compared to previous models like RNNs, LSTMs and DCNNs.

For computer vision applications, such as image classification, object detection, and semantic segmentation, the Transformer's ability to capture global dependencies across the entire image allows a more nuanced understanding and extraction of the features that exist in the image. Unlike DCNNs, which primarily focus on local features through convolutional filters, Transformers consider the full context of the image, enabling the model to make more informed predictions based on the global structure and relationships between different image regions [22]. This global perspective is particularly beneficial for tasks requiring detailed scene understanding and object interactions, illustrating the Transformers' potential to redefine approaches in computer vision. For instance, the prediction of the next frame from the current frame of video data requires both global and local understanding of the image contents and features.

There are plenty of ConvLSTM models that have been introduced to the literature using attention mechanism, such as [23–25]. We selected one of the most recent attention-based models, SwinLSTM for this study. We investigated the predictive capabilities of using LSTM models in predicting flashover in room fire incidents. Both ConvLSTM and SwinLSTM models were trained using our recorded thermal IR vision data from actual room fire tests for the task of flashover prediction. SwinLSTM [26] is an advanced combinatorial model that utilizes Swin Transformer blocks. Like ConvLSTM, SwinLSTM benefits from the power of DCNN and LSTM, while Swin Transformer blocks improve the prediction performance of the model considerably. The innovative approach of the self-attention mechanism of SwinLSTM with both enhanced global and local feature extraction and memorization drove the model to efficiently process high-dimensional video data collected from room fire tests with higher accuracy than previous models. Our experimental results of the SwinLSTM showcased the potential of the model to revolutionize fire incident prediction, offering insights that can lead to improved safety for firefighters and occupants. In the application for actual room fire flashover prediction, the compared performance of a ConvLSTM and SwinLSTM revealed the better prediction capability of the SwinLSTM.

2. Recurrent Neural Networks

Recurrent Neural Networks (RNNS) are developed to predict events in spatiotemporal data. In this section, we first look at the general idea of the ConvLSTM model, which is one RNN model and how convolutional operation and transformer mechanism have been introduced to these models for better performance and higher accuracy.

2.1. Convolutional Long Short-Term Memory (ConvLSTM)

Advancements in AI, particularly in deep learning, and the use of thermal IR and vision cameras, have equipped researchers with powerful tools for monitoring and detecting rapid fire and smoke growth phenomena, such as flashover [11,13,14,27,28]. Despite these advancements, the prediction of such rapid-fire growth events using vision-based data remains a challenge where the AI model should predict future video frames from limited past multi-dimensional spatiotemporal data. Traditional RNNs, capable of handling sequential and time-series one-dimensional data, face challenges with long-term dependencies due to their tendency to forget previous information. LSTM networks, introduced to overcome these limitations, incorporate mechanisms for selectively remembering and forgetting information, making them better suited for capturing long-term dependencies [20]. The structure of a standard LSTM module, including a forget gate, facilitates this by managing information flow across sequences, thus serving as a memory unit for predicting subsequent time steps.

While LSTM models excel at processing one-dimensional data, such as text and signals, they fall short in handling multi-dimensional spatial sequence data, like images and videos. ConvLSTM, an extension of the LSTM, integrates convolutional operations within the LSTM structure, making it suitable for analyzing spatial-temporal data like video data. This hybrid model allows for the direct processing of two-dimensional image data along with temporal sequence information, providing a more robust framework for spatial-temporal data analysis compared to the original LSTM network [21]. Detailed information about LSTM and ConvLSTM model structures is provided in the Appendix A. It is noteworthy to mention that the ConvLSTM module is different from the ConvLSTM network. The former is a module combination of components that act as a network layer, while the latter is an artificial neural network consisting of ConvLSTM layers and other types of network layers.

The ConvLSTM network architecture, illustrated in Figure 1, consists of multiple ConvLSTM modules arranged to process video data efficiently. To improve the network's training capability of ConvLSTM, a batch normalization layer is positioned between ConvLSTM modules. Input data passes through all ConvLSTM layers, batch normalization layers, and dropout layers. Finally, the input data passes through a 3D convolutional layer with a Sigmoid activation function. This architecture, detailed in Figure 1, demonstrates the network component design and their functionality used for predicting complex phenomena like flashover from IR thermal data. In general, benefiting from both temporal and spatial information captured in the sequence of image data, ConvLSTM predicts the next frame of the current video in the future.

2.1.1. Transformers

Transformers are composed of two main components: the encoder and the decoder. Each encoder layer processes input data with a self-attention mechanism before passing it through a feedforward neural network. The decoder follows a similar structure, but includes an additional step of attention over the encoder's output to make predictions [22]. The core idea behind Transformers is to model relationships between all parts of the input data, regardless of their positions. For instance, when applied to natural language processing applications, this allows the model to directly learn the relationship between distant words in a sentence, without processing the intermediate words sequentially. This capability comes from the self-attention mechanism, which computes attention scores representing the importance of every part of the input data to every other part.

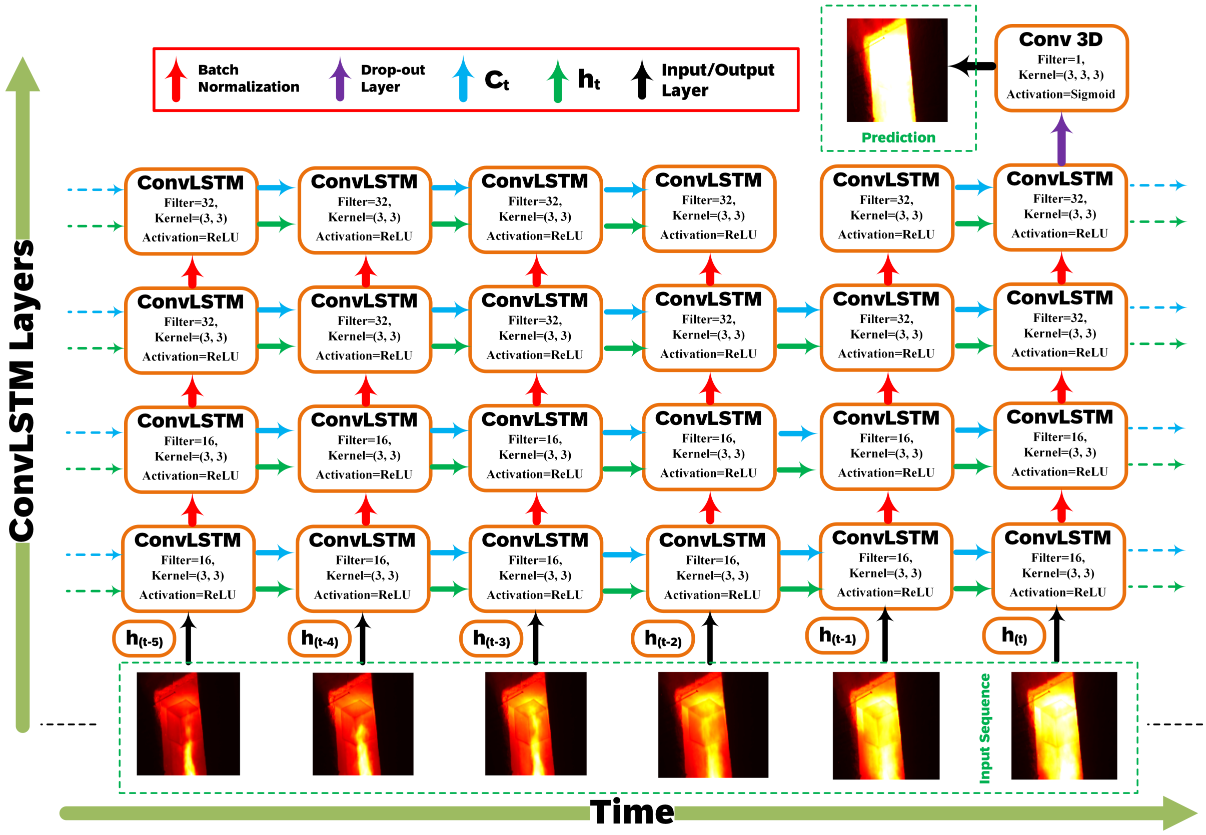

Figure 1. The architecture of the ConvLSTM network used in this study, key components of the network, and their functions.

Building upon their success in natural language processing, Transformers have been efficiently adapted to tackle complex tasks within the field of computer vision. This adaptation was marked by the introduction of Vision Transformer (ViT) [29], which applied the self-attention mechanism to image pixels or patches, treating them analogously to words in a sentence. As outlined in [29], ViT divides an image into fixed-size patches, linearly embeds each of them, adds positional embedding which helps the Transformer remember where each piece was in the original picture, and feeds the sequence of embedding into a standard Transformer encoder.

2.1.2. Swin Transformer

Swin Transformer is a variant of the Transformer architecture specifically adapted for computer vision tasks [30]. The Swin Transformer model builds upon the strengths of the original Transformer model by introducing a hierarchical structure that allows the model to efficiently process image data of varying size and complexity, making it particularly suitable for a wide range of computer vision tasks, such as image classification, object detection, and semantic segmentation [30]. The Swin Transformer processes images through a series of non-overlapping local windows. To capture global information, it alternates between these windows and “shifted” windows in subsequent layers. This shifting mechanism enables cross-window connections without significantly increasing computational complexity, thus efficiently capturing both local and global contextual information.

By limiting self-attention computation to within local windows and using relative position bias to maintain translation invariance, the Swin Transformer significantly reduces the computational cost associated with the self-attention mechanism. This efficiency makes it scalable to large images and dense prediction tasks. The hierarchical and modular nature of the Swin Transformer allows that to be easily integrated into existing computer vision architectures and scaled for various applications, from small-scale tasks to large, complex datasets and other deep learning models.

2.1.3. SwinLSTM

SwinLSTM offers a ground-breaking approach to spatiotemporal prediction by combining the self-attention capabilities of Swin Transformers with the sequential data processing strengths of LSTM networks. This hybrid model excels at capturing both extensive spatial dependencies and intricate temporal dynamics, positioning it as a superior choice for complex prediction tasks, such as rapid-fire spread anticipation [26]. The essence of SwinLSTM is encapsulated in its innovative cell design, which employs self-attention to process two-dimensional data, diverging significantly from the traditional convolutional methods utilized in predecessors like ConvLSTM. This architectural shift enables SwinLSTM to more effectively comprehend global spatial relationships within data, thereby boosting its predictive performance for spatiotemporal multi-dimensional data sequences.

Two variants of SwinLSTM have been proposed [26] to cater to different complexities of prediction tasks: SwinLSTM-B and SwinLSTM-D. SwinLSTM-B, the base model, incorporates a single SwinLSTM cell, emphasizing efficiency and simplicity while still delivering on the model's promise of effectively capturing spatial and temporal dependencies. On the other hand, SwinLSTM-D is designed as a deeper, more complex model with multiple SwinLSTM cells. It includes patch merging and expanding layers to handle spatiotemporal data at varying scales, making it particularly powerful at addressing more challenging spatiotemporal prediction tasks. For the fire growth prediction task, the power of SwinLSTM in discerning spatial layouts and temporal progressions offers a promising path toward enhancing predictive models. The integration of SwinLSTM into our predictive framework also enables us to extract the intricate dynamics of fire spread over time. The introduction of SwinLSTM-B and SwinLSTM-D variants handle a broad spectrum of application needs, from basic to complex spatiotemporal prediction challenges, exemplifying the versatility and robustness of this novel deep learning algorithm.

2.1.4. Model Architecture

The architecture of SwinLSTM is outlined in Figure 2, introducing two model variations: SwinLSTM-B and SwinLSTM-D. Figure 2a shows the fundamental SwinLSTM cell structure, whereas SwinLSTM cell integrates Swin Transformer blocks (STB) with a simplified LSTM structure around the one for enhanced spatiotemporal representation. The LP block in the structure denotes linear projection. Patch embedding is the process of dividing an image into smaller sections, transforming these sections into a series of vectors, and adding positional information to help a model understand the spatial relationships within the image. In SwinLSTM architectures, input images are passed through a patch embedding layer prior to SwinLSTM cells. This way, input images are transformed into sequences of patches to extract spatiotemporal features. Figure 2b illustrates the architecture of a SwinLSTM-B model that utilizes a single SwinLSTM cell. As explained, using SwinLSTM-B, the input image at the current time X_t is divided into patches, embeds these patches, and then processes them through the SwinLSTM cell alongside the previous time step's hidden and cell states (respectively, H_{t-1} and C_{t-1}). The output is then decoded by a reconstruction layer to predict the next frame. Figure 2c shows SwinLSTM-D architecture. It expands on the base model by incorporating multiple SwinLSTM cells and including Patch Merging and Patch Expanding layers for downsampling and upsampling, respectively. This design facilitates processing across different scales by using hidden and cell states (respectively, $H_{t-1}^{l=1,\ldots,m}$ and $C_{t-1}^{l=1,\ldots,m}$, which l denotes the number of the layer) of previous time steps and layers outputs, enhancing the model's ability to capture complex spatiotemporal dynamics. SwinLSTM-D follows a similar process to SwinLSTM-B, but with added complexity for handling multi-scale spatiotemporal information.

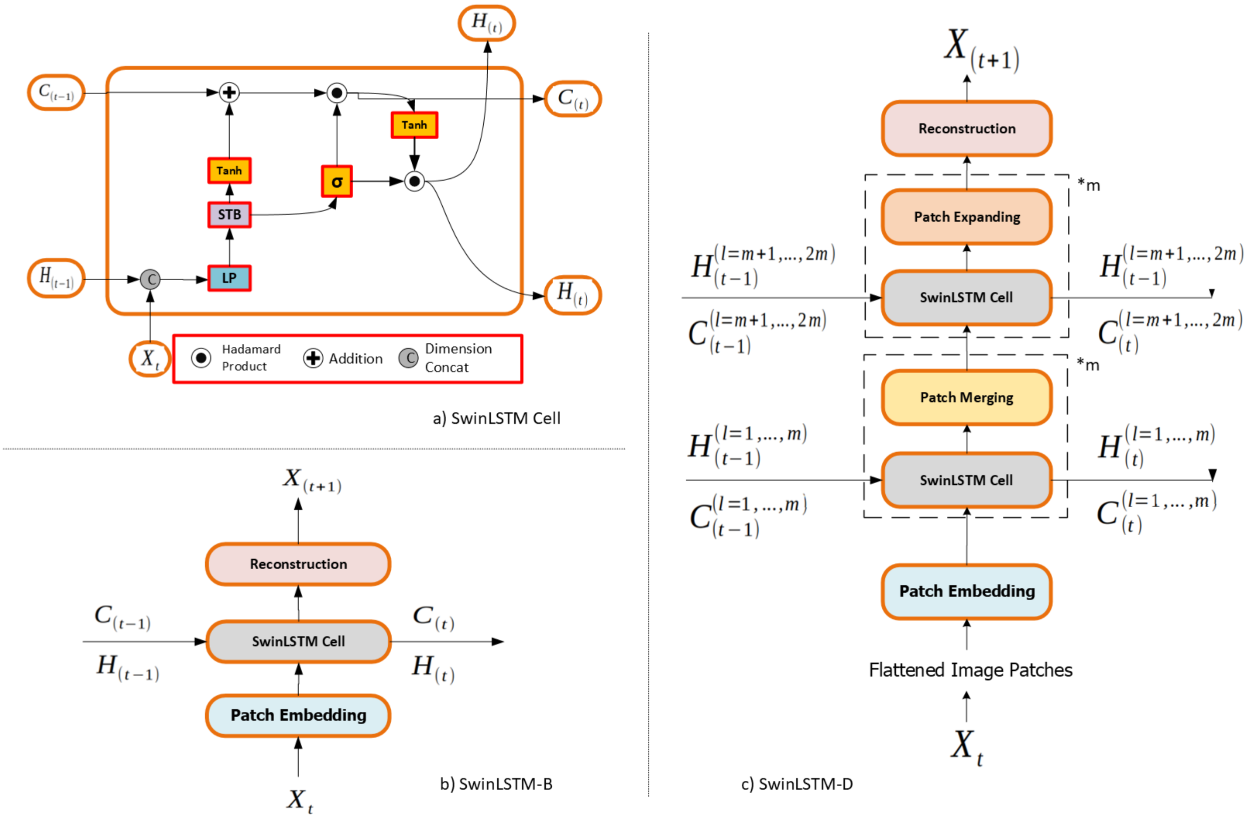

Figure 2. (**a**) the architecture of the SwinLSTM recurrent cell. STB and LP denote Swin Transformer blocks and Linear Projection; (**b**) the architecture of the SwinLSTM-B model, which contains a single SwinLSTM cell; and (**c**) the architecture of the deeper version of the SwinLSTM with multiple SwinLSTM cells, named SwinLSTM-D.

3. Results

3.1. Data Collection and Preparation

To train LSTM-based deep learning models for this study, we created various video datasets using thermal infrared (IR) video frames recorded by a FLIR T650sc camera during several full-scale room fire test experiments [2]. The details of the room set-ups and fire tests are provided in [2]. IR video data were first re-scaled into smaller sizes: 16 × 16, 32 × 32, 64 × 64, and 128 × 128. However, following the architecture of the original SwinLSTM work [26], we used videos with the size of 64 × 64 in our test experiments. After re-scaling and calculating the frame rate of raw IR videos captured from room fire test experiments, we retained one frame per second for generating training and testing videos while ensuring that all critical events were included in our datasets. We normalized the entire dataset using Min-Max normalization, and then separated the dataset into small video batches. The new datasets are tensors containing short-duration videos (i.e., 20 s, 30 s, 50 s, and 100 s in length). In this way, we made 16 various spatial and temporal datasets (with different scales and time durations). Again, following the common video size in the literature [21,26], we selected the 20 s dataset for training and testing in this study. In general, the dataset we used for training and testing of our deep learning models had a tensor of length 3338 × 20 × 64 × 64 × 1. We randomly separated the dataset into 70%, 20%, and 10% for training, validation, and test purposes, respectively. Figure 3 shows two random sample data from our generated dataset. Note that the frames selected every 5 s mean that we are looking at a video of size 95 s. We did this to increase the variation between frames of videos in order to force models to learn images with higher amounts of variations, resulting in better model generalization.

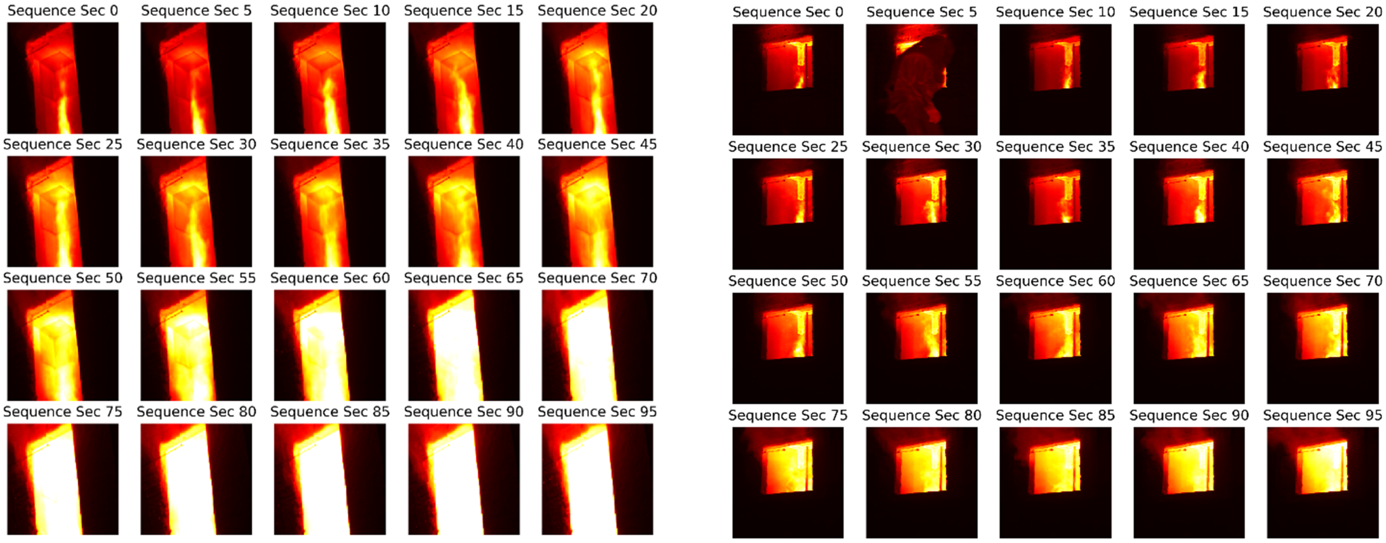

Figure 3. Two randomly selected data samples from the entire dataset. For the sake of illustration, we showed every 5 s. Red and yellow colors are lower and higher range of temperatures, respectively.

3.2. SwinLSTM Implementation and Setup

In order to train the LSTM-based models for evaluation and comparison, we implemented each model using the PyTorch library. We trained models for 10 epochs with a batch size of 16 using one NVIDIA Tesla K80 and 16 GB of Memory. The Adam Optimization [31] is used as the optimizer with its default parameters utilizing binary cross-entropy as the loss function. As a random illustration of one training trend of SwinLSTM architecture, Figure 4 shows that SwinLSTM architecture almost converged after a few epochs using our experimental setup regarding the trend of mean square error (MSE) as the criteria for training performance.

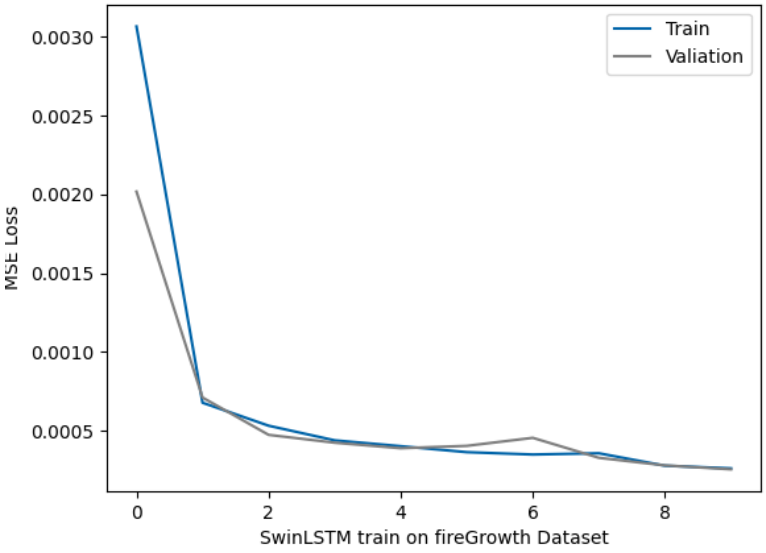

Figure 4. Training and validation of SwinLSTM for room fire IR data.

We used SwinLSTM-D architecture for the evaluation in this study, also to report the results from SwinLSTM-B, briefly. Therefore, for simplification from hereon, we use SwinLSTM interchangeably with the SwinLSTM-D. The pre-trained weights from the training of the SwinLSTM model on the Moving-MNIST dataset were used as the initial training weights of the model. The Moving-MNIST dataset is a challenging benchmark designed for evaluating image sequence prediction models. It consists of sequences of frames showing two digits moving inside a 64 × 64 grid, simulating complex, dynamic patterns that require advanced temporal and spatial understanding for accurate prediction. This pre-training provided our model with a robust foundational knowledge of handling dynamic sequences, which is crucial for understanding the intricate movements within fire-related video datasets. Table 1 illustrates the comparison size of different LSTM-based models used in this study. Looking at the table, it is clear that SwinLSTM achieved a better MSE loss value (definition of MSE provided in Equation (3)) with the cost of slower performance in comparison to the other two models.

Table 1. Comparison table for the number of parameters used in each LSTM-based deep learning model. Latency was calculated by the average of the time needed for generating frames in different test videos.

Methods	ConvLSTM	SwinLSTM
Parameters	3.8 Mb	20.1 Mb
Latency	27 ms ± 0.12	35 ms ± 0.65

3.3. Qualitative Results

We applied the trained SwinLSTM deep learning models for the test datasets that we collected and created, containing 338 fire content video IR data from different room fire test experiments. In Figure 5, the top row is ten past seconds, and the second row is ten seconds of the future (i.e., ground truth frames). The third row is the prediction results of SwinLSTM for ten seconds in the future. From the figure, it is clear that SwinLSTM is capable of generating future frames that include fine details and subtle changes in the image. For instance, the location of the window was not fixed in the experiments, but the model can predict that accurately. It is observed that the loss of details from the future frames is inevitable for the three models; however, SwinLSTM generates future frames with more details regarding the flame heights compared to the ground truth frames.

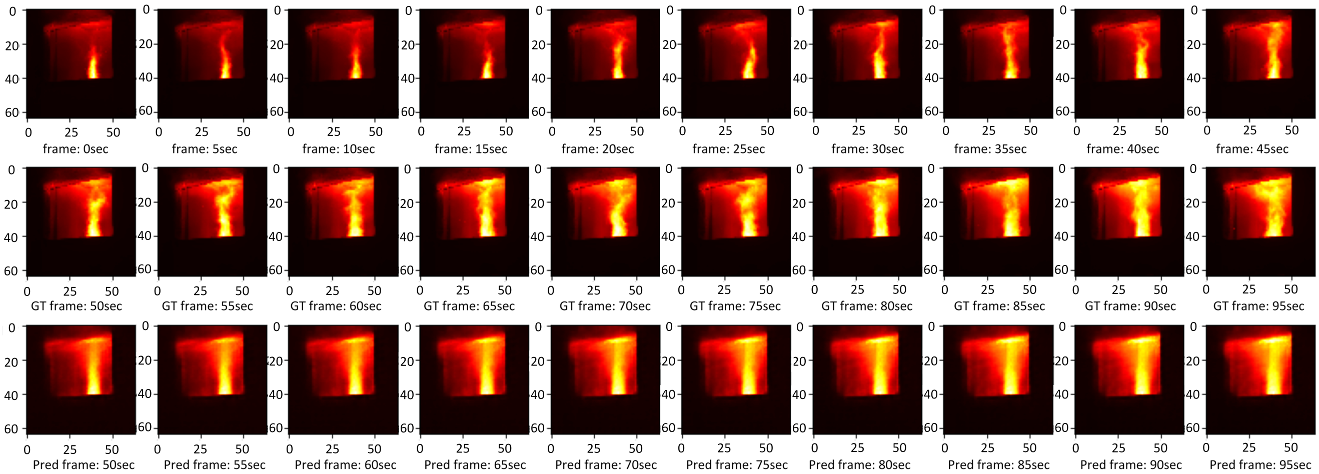

Figure 5. Results of SwinLSTM applied on test dataset.

In Figure 6, there are randomly selected results of the trained ConvLSTM used in our study applied for our test IR video dataset. The results show that although ConvLSTM is good at predicting future frames, it cannot predict details of the fire dynamics (e.g., flame movement details) in the image with high precision. In general, LSTM-based models are able to provide future frames by keeping the most of abstract content such as the location of the opening and the general shape of the flame. For a better comparison, we need to look at the quantitative results. To show that the temperature content in images is also approximately retained in generated future frames, we also provided the colormap of images in Figure 6. It is difficult to compare between temperature of ground truth frames and the generated ones since conversion from RGB values to temperature data comes with the loss of information in this study. Approximately, earlier generated future frames have more correlation with ground truth frames than the further future generated frames. However, in total, the range of temperature is reasonable and acceptable.

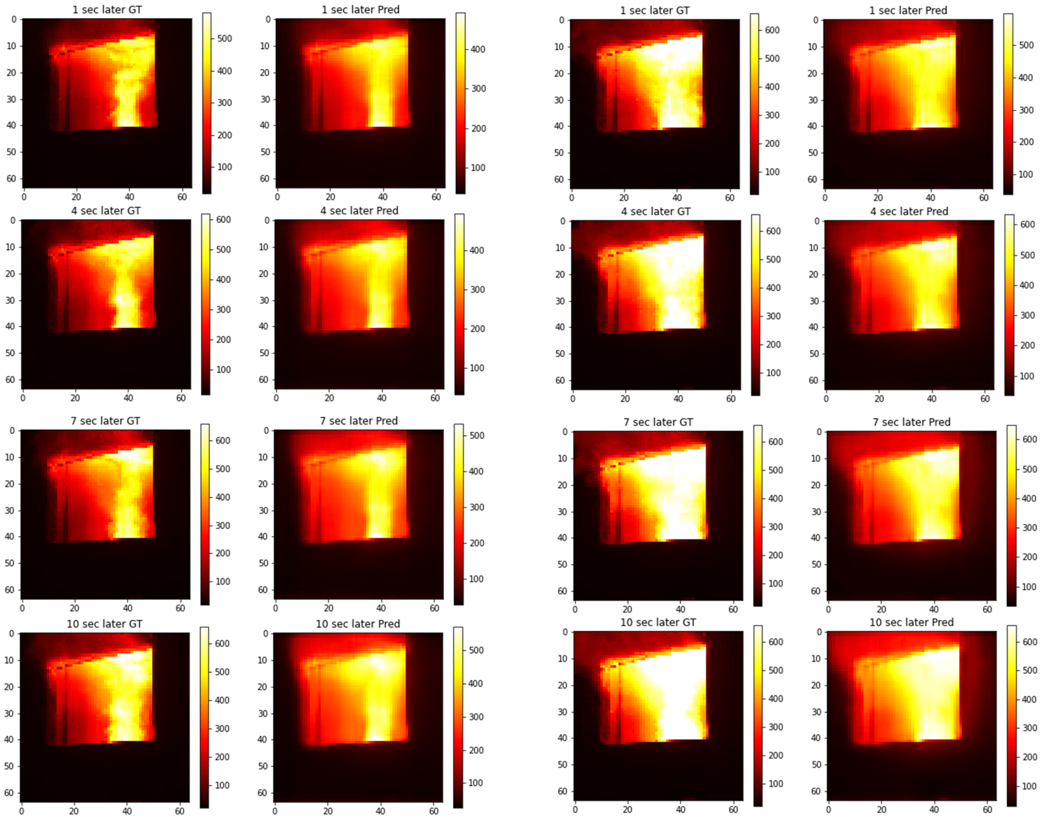

Figure 6. Selected frames from test dataset for better illustration of prediction details by ConvLSTM model. The temperature colormap also provided for better comparison between ground truth images and the predicted images.

3.4. Quantitative Results

Although SwinLSTM could predict fire growth images similar to reality, it is important to know how far the prediction of the model deviates from the ground truth images. For this reason, we are reporting here the results of common quantitative assessment measures. The structural similarity index measure (SSIM) is a method for predicting the perceived quality of digital images. It is also used for measuring the similarity between two images. The SSIM between predicted and ground truth frames is defined as the mathematical expression in Equation (1). SSIM values close to value one indicate a high degree of similarity.

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{1}$$

where x and y are the compared prediction and ground truth images, μ_x and μ_y represent the average intensities of x and y images, σ_x^2 and σ_y^2 represent the variance of x and y, σ_{xy} is the covariance of x and y, and C_1 and C_2 are constant. We also reported MSE and Mean Absolute Error (*MAE*) as other metrics between predicted and ground truth frames (see Equations (2) and (3)). A simple explanation for these metrics is that *MAE* and *MSE* measure the average and variance of the residuals between predicted and ground truth images, respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

where y_i and $\hat{y}_i$ are image pixels at location i of the predicted and ground truth images, and n is the spatial dimension of the images. The results of the comparison assessment measure for one test video sample (qualitatively illustrated in Figure 5) used for testing SwinLSTM are illustrated in Figure 7. The fire details, such as flame area, were not clear in the qualitative comparison figure, but here, from Figure 7, it can be seen that the value of SSIM is decreasing while the value of MAE is increasing. The reason is that the more time passes in the future, the more difficult it is for the model to predict the future. However, the difference between the values for the next frame in the future and 10 frames in the future is not significant. The MSE trend value was constant while the MAE increased. This can also be expected since the average value of intensities in the frames is similar over time while the variance of average intensity per frame between ground truth images and prediction increases due to the loss of details by passing the time in the future. Again the differences are meaningful, but not significant, and in general SwinLSTM can predict images with high SSIM values.

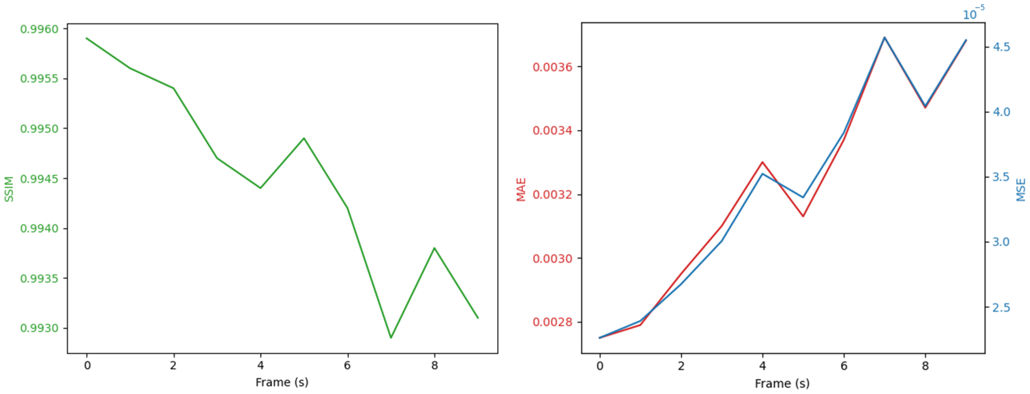

Figure 7. Quantitative comparison analysis of ground truth image and prediction result image of SwinLSTM applied on IR video data shown in Figure 5.

Table 2 shows a notable performance comparison of LSTM-based models applied to our test IR video dataset. The pre-trained weights from the Moving-MNIST dataset boosted LSTM-based models with a pre-understanding of motion, enabling it to perform more effectively to the specific challenges presented by the image sequence of fire dynamics. From Table 2, the average SSIM values for the results of almost all LSTM-based models are around 1, which confirms the high similarity between ground truth images and predicted images.

Table 2. The performance (average and standard deviation) comparison table of LSTM-based deep learning models on test IR video data.

Model	**ConvLSTM**	**SwinLSTM-B**	**SwinLSTM-D**
SSIM	0.97 ± 0.062	0.96 ± 0.021	0.98 ± 0.010

To further evaluate the performance of the SwinLSTM model, we calculated the same similarity and error measurements for a few novel test videos data collected from a completely different room fire test IR dataset [32]. The graph in Figure 8 shows frame prediction results by SwinLSTM model for 20 randomly selected test videos from that dataset. It should be noted that these videos were unseen, and with almost different test setups (e.g., various opening sizes and locations, angle of view, temperatures, and fire behavior). As can be seen, the maximum range of error and similarity differences are still negligible, and the performance of the model for various data situations is significantly stable.

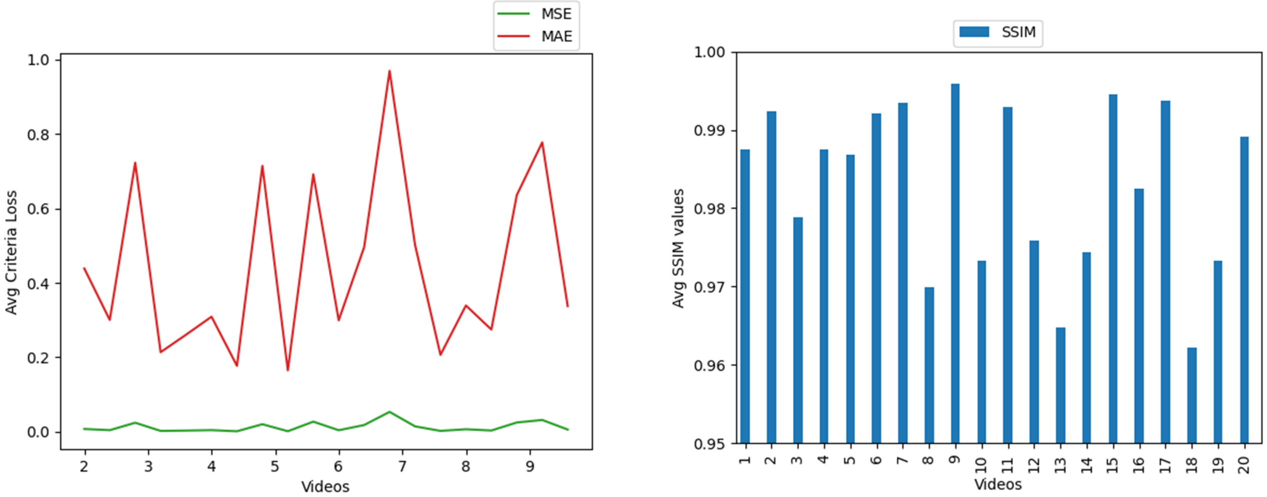

Figure 8. Performance of SwinLSTM in prediction of one frame in future for 20 randomly selected room fire test videos.

4. Discussion and Conclusions

In this study, we applied two deep learning architectures, SwinLSTM and ConvLSTM models, to tackle the challenge of predicting rapid room fire growth. The nature of fire growth in enclosed spaces is alarmingly rapid, leaving minimal time for occupants and firefighters to respond effectively. Furthermore, the potential for a fire to transition to a flashover stage, where it consumes a room entirely, underscores the urgency of accurately forecasting fire behavior to safeguard lives.

Traditionally, prediction efforts have relied on sensor-based data, requiring the prior installation of sensors and gauges in a room. While this approach helped establish the scientific understanding of flashover phenomena, it falls short in practical fireground scenarios due to its inherent limitations and lack of flexibility. Vision-based methods using thermal IR and RGB cameras have emerged as promising alternatives, overcoming some constraints of conventional techniques by enabling fire detection, location, and growth monitoring through dynamic visual data.

This study represents a significant advancement in fire behavior prediction. We harnessed the power of deep neural networks, specifically the SwinLSTM and ConvLSTM models, to leverage spatiotemporal data, a previously untapped strategy in this context. The adoption of Swin Transformers in our models is particularly noteworthy for its ability to capture complex spatial relationships and temporal dynamics in high-dimensional data. This approach allows for a nuanced understanding of fire development, offering the potential to predict critical fire behaviors, such as flashover, several seconds in advance with remarkable accuracy, instilling hope for improved fire safety. The findings demonstrate that Convolutional LSTM-based models such as ConvLSTM and SwinLSTM are capable of predicting general and abstract contents of future frames with acceptable detail contents useful to be employed in fire safety applications such as flashover detection, where the need is to predict the onset of flashover.

Based on the best of our knowledge from the literature, this study is the first sample of vision-based prediction of rapid-fire growth prediction. Our work in the use of convolutional LSTM-based deep learning models provides a promising framework for the further development of methods to predict and monitor fire growth in room fire incidents using vision-based data. In terms of future work, training these models on vision-based datasets with various content variations will improve the accuracy of the methods as well as predict frames farther into the future to save more lives and properties from room fires.

Author Contributions: Conceptualization, M.H.M.; methodology, M.H.M.; software, Y.L. and N.H.; validation, M.H.M. and Y.K.; data curation, M.H.M. and Y.L.; writing—original draft preparation, M.H.M.; writing—review and editing, M.H.M. and N.H.; visualization, M.H.M. and N.H.; supervision, Y.K.; project administration, Y.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data available on request due to restrictions e.g., privacy or ethical.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Basics of Recurrent Network Models

In this Appendix, we provided the fundamentals and details of recurrent and LSTM-based convolutional neural networks as a reference for non-expert reader. We recommend to review the appendix before reading the entire study for better understanding the methodology. Details of components in a recurrent cell can be seen in Figure A1C. From the figure, the output of a recurrent layer (see Figure A1A) is fed to the input in a loop. In reality, this small recurrent network with one layer can be represented in an unrolled version (Figure A1B). The mathematical expressions of one standard recurrent cell can be expressed by Equation (A1).

$$\begin{cases} h_t = \sigma(W_h \cdot h_{t-1} + W_x \cdot x_t + b), \\ \hfill y_t = h_t \end{cases} \tag{A1}$$

where x_t, h_t, and y_t denote the input vector, hypothesis vector or recurrent information, and the output vector of the recurrent cell at time t, respectively. W_h, W_x, and b are the network weights and the bias. If we omit $W_h \cdot h_{t-1}$ term in Equation (A1), then the cell is converted into one standard neural network. In neural networks, activation functions [33], here Sigmoid function, are used to give non-linearity to the network equations.

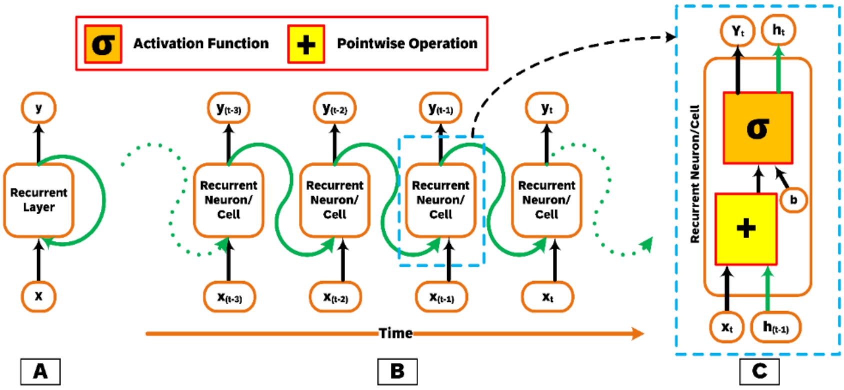

Figure A1. Schematic of (**A**) a simple standard recurrent layer, (**B**) unrolled version of a simple recurrent layer, (**C**) detailed components of a recurrent neuron or cell.

As explained in the introduction section, LSTM architecture is designed based on RNN models to alleviate the problem of memory loss for longer data sequences. The mathematical expressions of the LSTM with forget gate is presented by Equation (A2).

$$\begin{cases} f_t = \sigma(W_{fh} \cdot h_{t-1} + W_{fx} \cdot x_t + b_f), \\ i_t = \sigma(W_{ih} \cdot h_{t-1} + W_{ix} \cdot x_t + b_i), \\ c'_t = tanh(W_{c'h} \cdot h_{t-1} + W_{c'x} \cdot x_t + b_{c'}), \\ c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t, \\ o_t = \sigma(W_{oh} \cdot h_{t-1} + W_{ox} \cdot x_t + b_o), \\ h_t = o_t \cdot \tanh(c_t) \end{cases} \tag{A2}$$

Figure A2 is the implementation and deployment of Equation (A2) as a schematic for one LSTM cell or layer. The input data in a form of one-dimensional sequence is fed to the LSTM cell as X_t and states of previous layers h_{t-1} and C_{t-1} provided separately as inputs. The output of the cell will be the states of the cell in the current time sending to the next layer, as well as the hypothesis of the current layers h_t, which can be converted into the prediction y using a decoder layer (i.e., use of a loss function).

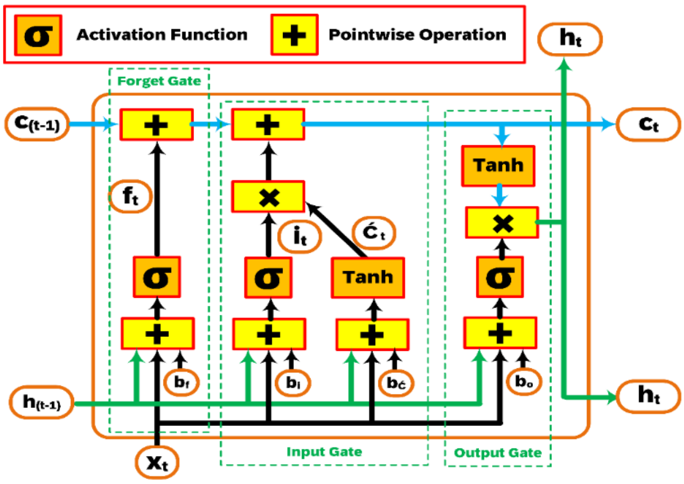

Figure A2. Schematic of original LSTM architecture with forget gate.

The two-dimensional version of LSTM cell was designed using convolutional operation. Mathematically, one ConvLSTM cell can be expressed as Equation (A3).

$$\begin{cases} f_t = \sigma(W_{fh} * h_{t-1} + W_{fx} * x_t + b_f), \\ i_t = \sigma(W_{ih} * h_{t-1} + W_{ix} * x_t + b_i), \\ c'_t = tanh(W_{c'h} * h_{t-1} + W_{c'x} * x_t + b_{c'}), \\ c_t = f_t \cdot c_{t-1} + i_t \cdot c'_t, \\ o_t = \sigma(W_{oh} * h_{t-1} + W_{ox} * x_t + b_o), \\ h_t = o_t \cdot \tanh(c_t) \end{cases} \tag{A3}$$

In a ConvLSTM cell, matrix multiplication at each gate of classical LSTM is replaced with convolution operation. Comparing Equation (A2) with Equation (A3), the only difference is the utilization of convolutional operation '$*$' instead of multiplication '$\cdot$' in gate sub-layers. In Figure A3, you can see the structure of one ConvLSTM layer.

Having ConvLSTM cells, it is possible to connect several ConvLSTM cells in various configurations and with other types of convolutional layers (e.g., batch normalization) to create a complete ConvLSTM network for the task of video frame prediction.

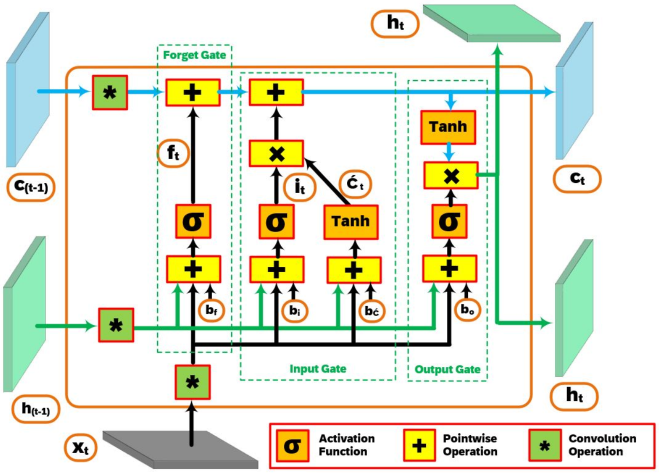

Figure A3. Details of the recurrent structure of a ConvLSTM module.

References

1. Statistics Canada. *Incident-Based Fire Statistics, by Type of Fire Incident and Type of Structure*; Statistics Canada: Ottawa, ON, Canada, 2017. [CrossRef]
2. Mozaffari, M.H.; Li, Y.; Ko, Y. Real-time detection and forecast of flashovers by the visual room fire features using deep convolutional neural networks. *J. Build. Eng.* **2023**, *64*, 105674. [CrossRef]
3. Peacock, R.D.; Reneke, P.A.; Bukowski, R.W.; Babrauskas, V. Defining flashover for fire hazard calculations. *Fire Saf. J.* **1999**, *32*, 331–345. [CrossRef]
4. Cortés, D.; Gil, D.; Azorín, J. Fire Science Living Lab for Flashover Prediction. *Proceedings* **2019**, *31*, 87. [CrossRef]
5. Mozaffari, M.H.; Li, Y.; Weinfurter, M.; Ko, Y. Study of flashover in full-scale room fires using imaging technologies. In *Technical Report*; National Research Council of Canada: Statistics Canada: Ottawa, ON, Canada, 2024. [CrossRef]
6. Kim, H.J.; Lilley, D.G. Flashover: A study of parameter effects on time to reach flashover conditions. *J. Propuls. Power* **2002**, *18*, 669–673. [CrossRef]
7. Zhang, Y.; Wang, L. Research on flashover prediction method of large-space timber structures in a fire. *Materials* **2021**, *14*, 5515. [CrossRef] [PubMed]
8. Tam, W.C.; Fu, E.Y.; Peacock, R.; Reneke, P.; Wang, J.; Li, J.; Cleary, T. Generating synthetic sensor data to facilitate machine learning paradigm for prediction of building fire hazard. *Fire Technol.* **2020**, *59*, 1–22 . [CrossRef] [PubMed]
9. Huyen, A.; Yun, K.; De Baun, S.; Wiggins, S.; Bustos, J.; Lu, T.; Chow, E. Dynamic fire and smoke detection and classification for flashover prediction. In Proceedings of the Pattern Recognition and Tracking XXXII, Online, 12–16 April 2021; SPIE: Bellingham, WA, USA, 2021; Volume 11735, p. 1173502. [CrossRef]
10. Mozaffari, M.H.; Li, Y.; Ko, Y. Generative AI for Fire Safety. In *Applications of Generative AI*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 577–600. [CrossRef]
11. Mozaffari, M.H.; Li, Y.; Ko, Y. Detecting Flashover in a Room Fire based on the Sequence of Thermal Infrared Images using Convolutional Neural Networks. In Proceedings of the Canadian AI, Toronto, ON, Canada, 30 May–3 June 2022. [CrossRef]
12. Ko, Y.; Hamed Mozaffari, M.; Li, Y. Fire and smoke image recognition. In *Intelligent Building Fire Safety and Smart Firefighting*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 305–333. [CrossRef]
13. Yun, K.; Bustos, J.; Lu, T. Predicting rapid fire growth (flashover) using conditional generative adversarial networks. *arXiv* **2018**, arXiv:1801.09804. [CrossRef]
14. Francis, J.; Chen, A. Observable characteristics of flashover. *Fire Saf. J.* **2012**, *51*, 42–52. [CrossRef]
15. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
16. Kim, B.; Lee, J. A video-based fire detection using deep learning models. *Appl. Sci.* **2019**, *9*, 2862. . [CrossRef]
17. Han, D.; Lee, B. Flame and smoke detection method for early real-time detection of a tunnel fire. *Fire Saf. J.* **2009**, *44*, 951–961. [CrossRef]
18. Zhou, Y.; Dong, H.; El Saddik, A. Deep learning in next-frame prediction: A benchmark review. *IEEE Access* **2020**, *8*, 69273–69283. [CrossRef]
19. Mikolov, T.; Karafiát, M.; Burget, L.; Cernockỳ, J.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the Interspeech, Makuhari, Chiba, Japan, 26–30 September 2010; Volume 2, pp. 1045–1048.
20. Hochreiter, S. *Long Short-Term Memory*; Neural Computation MIT-Press: La Jolla, CA, USA, 1997.
21. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810.
22. Vaswani, A. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, 6000–6010. [CrossRef]

23. Li, D.; Chen, Q. Deep reinforced attention learning for quality-aware visual recognition. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part XVI 16, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 493–509. [CrossRef]
24. Huang, Z.; Liang, S.; Liang, M.; Yang, H. Dianet: Dense-and-implicit attention network. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 4206–4214. [CrossRef]
25. Lin, Z.; Li, M.; Zheng, Z.; Cheng, Y.; Yuan, C. Self-attention convlstm for spatiotemporal prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11531–11538. [CrossRef]
26. Tang, S.; Li, C.; Zhang, P.; Tang, R. Swinlstm: Improving spatiotemporal prediction accuracy using swin transformer and lstm. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 13470–13479.
27. Ajith, M.; Martínez-Ramón, M. Unsupervised segmentation of fire and smoke from infra-red videos. *IEEE Access* **2019**, *7*, 182381–182394. [CrossRef]
28. Muhammad, K.; Ahmad, J.; Mehmood, I.; Rho, S.; Baik, S.W. Convolutional neural networks based fire detection in surveillance videos. *IEEE Access* **2018**, *6*, 18174–18183. [CrossRef]
29. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26–30 April 2020.
30. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
31. Kingma, D.P. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
32. Bwalya, A.; Gibbs, E.; Lougheed, G.; Kashef, A. *Characterization of Fires in Multi-Suite Residential Dwellings: Final Project Report: Part 1—A Compilation of Post-Flashover Room Fire Test Data*; National Research Council of Canada: Ottawa, ON, Canada, 2014. [CrossRef]
33. Hayou, S.; Doucet, A.; Rousseau, J. On the Impact of the Activation function on Deep Neural Networks Training. In Proceedings of the 36th International Conference on Machine Learning. PMLR, Long Beach, CA, USA, 9–15 June 2019; Volume 97, pp. 2672–2680.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.