# MCSDNet: Mesoscale Convective System Detection Network via Multi-scale Spatiotemporal Information

Jiajun Liang, Baoquan Zhang, Chuyao Luo, Xutao Li, Yunming Ye, Xukai Fu

***Abstract*—The accurate detection of Mesoscale Convective Systems (MCS) is crucial for meteorological monitoring due to their potential to cause significant destruction through severe weather phenomena such as hail, thunderstorms, and heavy rainfall. However, the existing methods for MCS detection mostly targets on single-frame detection, which just considers the static characteristics and ignores the temporal evolution in the life cycle of MCS. In this paper, we propose a novel encoder-decoder neural network for MCS detection(MCSDNet). MCSDNet has a simple architecture and is easy to expand. Different from the previous models, MCSDNet targets on multi-frames detection and leverages multi-scale spatiotemporal information for the detection of MCS regions in remote sensing imagery(RSI). As far as we know, it is the first work to utilize multi-scale spatiotemporal information to detect MCS regions. Firstly, we design a multi-scale spatiotemporal information module to extract multi-level semantic from different encoder levels, which makes our models can extract more detail spatiotemporal features. Secondly, a Spatiotemporal Mix Unit(STMU) is introduced to MCSDNet to capture both intra-frame features and inter-frame correlations, which is a scalable module and can be replaced by other spatiotemporal module, *e.g.*, CNN, RNN, Transformer and our proposed Dual Spatiotemporal Attention(DSTA). This means that the future works about spatiotemporal modules can be easily integrated to our model. Finally, we present MCSRSI, the first publicly available dataset for multi-frames MCS detection based on visible channel images from the FY-4A satellite. We also conduct several experiments on MCSRSI and find that our proposed MCSDNet achieve the best performance on MCS detection task when comparing to other baseline methods. Particularly, MCSDNet shows remarkable capability in extreme conditions where MCS regions are densely distributed. We hope that the combination of our open-access dataset and promising results will encourage the future research for MCS detection task and provide a robust framework for related tasks in atmospheric science. Our code is available [1].**

***Index Terms*—mesoscale convective system detection, semantic segmentation, multi-scale spatiotemporal information, dual spatiotemporal attention, encoder-decoder neural network**

## I. Introduction

CONVECTIVE weather is a type of sudden and intense meteorological disaster that can cause significant destruction. It is often accompanied by severe weather phenomena such as hail, thunderstorms, strong winds, and short-term heavy rainfall, posing a serious threat to the life and property safety of the public. Cloud masses generated by convection are commonly referred to as convective clouds. The air inside these clouds is in an extremely unstable state, characterized by high temperatures in the lower atmosphere and low temperatures in the upper atmosphere. Convective weather, with its short occurrence time, sudden intensity, and significant impact, has always been a challenging aspect in meteorological monitoring. Convective weather is caused by mesoscale convective system(MCS), which is a cloud cluster with a convective core and extends approximately 100 km in a direction, forming a general precipitation area [1]. The key of convective weather detection is how to detect MCS regions. With the advancement of remote sensing satellite technology, researchers have attempted to utilize remote sensing imagery(RSI) to detect MCS [2] [3]. However, MCS detection faces the issue of class imbalance, where the number of negative samples far exceeds the number of positive samples. As shown in Fig. 1, among the observation data collected in China in 2018, more than 60% of satellite images contained only 1% of MCS regions. Additionally, MCS has a complex and dynamic life cycle, which consists of three stages: initiation, maturity, and a decaying phase where the location and shape of convective cloud change over time. When performing detection of MCS regions, it is necessary to consider both the spatial and temporal information. MCS detection is a challenging issue with few existing solutions but is worth to be addressed.

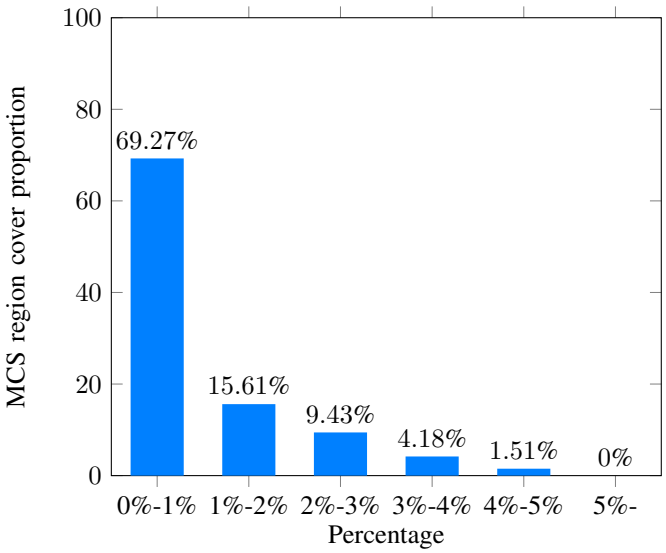



Fig. 1: The proportion distribution of MCS regions among the observation data collected in China in 2018.

Jiajun Liang, Baoquan Zhang, Chuyao Luo, Xutao Li, and Yunming Ye are with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, Guangdong, China; Xukai FU is with Beijing Jinkai New Energy Environmental Technology Co., Ltd.

E-mail: liangjiajun2002@163.com, baoquanzhang@hit.edu.cn, {lixutao, yeyunming}@hit.edu.cn, luochuyao@stu.hit.edu.cn, 2510442827@qq.com

Corresponding author: Baoquan Zhang.

[1]https://github.com/250HandsomeLiang/MCSDNet.git

In terms of MCS detection, fixed brightness temperature threshold method(BT) is the most commonly used detection technique. It distinguishes MCS regions from other objects at the pixel level by utilizing their spectral and wavelength differences [4]. However, the BT threshold is not universally and it can produce significant detection differences due to varying conditions such as geographical location, temperature, and properties of the lower atmosphere.

In recent years, researchers have tried to apply machine learning model to atmospheric science [5] [6]. Different from BT threshold, the machine learning model doesn't need to understand complex atmospheric physical and dynamic processes and it is a data-driving method. However, MCS detection algorithms based on machine learning are not end-to-end solutions. Instead, their detection effectiveness heavily relies on the quality of feature engineering, resulting in insufficient intelligent decision-making capabilities [7].

Benefiting from the development of deep learning, end-to-end models based on neural network have been applied to RSI analysis, *e.g.*, hyperspectral image classification [8] [9] [10], cloud detection [11] [12] [13], precipitation nowcasting [14] [15] and so on. However, these models are not designed for MCS detection and they don't take into account the dramatic changes of MCS in the whole life cycle. Besides, the existing models target on single-frame image and don't consider the temporal evolution in the life cycle of MCS. It is challenging to detect MCS regions accurately.

Similar to [16] [17], MCS detection task can be regarded as a semantic segmentation task in computer vision. Given a satellite image, we aim to use a specific algorithm to classify each pixel in the original image . Since each pixel in the original image can only belong to either the category of MCS region or not, the task of MCS detection can also be considered as a pixel-level binary classification task. We can utilize the efficient models in semantic segmentation task to detect MCS regions.

These methods mentioned above all target on single-frame detection and ignore the temporal evolution in the life cycle of MCS. Considering the dramatic and changeable characteristics of MCS, they are limited to achieve excellent performance. In this paper, we argue the significance of spatiotemporal information and propose an encoder-decoder neural network for MCS detection(MCSDNet). Different from the existing models, MCSDNet targets on multi-frames MCS detection and takes image sequence as input. This model can capture the temporal evolution of MCS by exploring cross-frames variations in image sequence. In MCSDNet, we capture multi-scale spatiotemporal information by incorporating the feature maps from different encoder levels. Then, we apply a Spatialtemporal Mix Unit(STMU) to explore the relation and consistency between consecutive frames. STMU is a scalable spatiotemporal module to capture temporal evolution of MCS. In our implementation, we design a new spatiotemporal module named Dual Spatiotemporal Attention(DSTA) as STMU. Different from previous spatiotemporal modules [18] [19] [20], DSTA can effectively capture both intra-frame features and inter-frame correlations. Compared to other models, MCSDNet considers both spatial and temporal information of MCS regions and gains better performance in MCS detection task.

Our main contributions can be summarized as follows:

- We propose an encoder-decoder neural network for MCS detection(MCSDNet). MCSDNet has a simple architecture and is easy to expand. The model utilizes a Spatiotemporal Mix Unit(STMU) to capture temporal evolution of MCS regions, which is scalable modules and can be replaced by other spatiotemporal modules, *e.g.*, CNN, RNN, Transformer and our proposed Dual Spatiotemporal Attention. Besides, MCSDNet maintains the original performance standards of the spatiotemporal modules without any compromise. This means that the future works about spatiotemporal modules can be easily incorporated into our model.
- Different from the existing methods, our model targets on multi-frames MCS detection and captures the temporal evolution in the life cycle of MCS by exploring the cross-frames correlations among image sequence. In MCSDNet, we design a multi-scale spatiotemporal information module to extract multi-scale semantic from different encoder levels, which makes out model more suitable to extreme conditions. As far as we know, it is the first work to utilize multi-scale spatiotemporal information to detect MCS regions.
- We create and publicly share a Mesoscale Convective System Detection dataset based on the captured images from FY-4A satellite: Mesoscale Convective System Remote Sensing Image(MCSRSI). As far as we know, it is the first publicly available dataset for multi-frames MCS detection task based on visible channel images. Besides, we conduct experiments on it. The results demonstrate that MCSDNet can achieve state-of-the-art performance on MCS detection task. It provides a baseline for future research in this direction.

The rest of this work is organized as follows: In Section II, we briefly review related works on MCS detection. Section III describes our model(MCSDNet) in details. In Section IV, we introduce the newly created dataset for MCS detection: MCSRSI and present the experimental results on it. Finally, the conclusion is summarized in Section V.

## II. Related Work

### A. Mesoscale Convective System Detection Methods

When detecting MCS regions, threshold methods is the most commonly used methods [21], [22], [23], [24]. Brightness temperature(BT) method utilizes the relatively low cloud top temperature in convective clusters to detect MCS. The threshold is usually ranging from 208 K to 255 K [25] [26]. BT difference(BTD) is based on water vapor sensitivity and uses the brightness temperature difference to explore convective clouds. Different from threshold methods, machine learning methods do not need to understand complex atmospheric physical and spectral channel information. It is gradually applied to MCS detection. Kim *et al.* [27] utilize the images from the Fengyun-8 satellite to construct random forest and logistic regression models which can effectively locate MCS regions. Yang *et al.* [3] construct MCS detection and tracking algorithms for

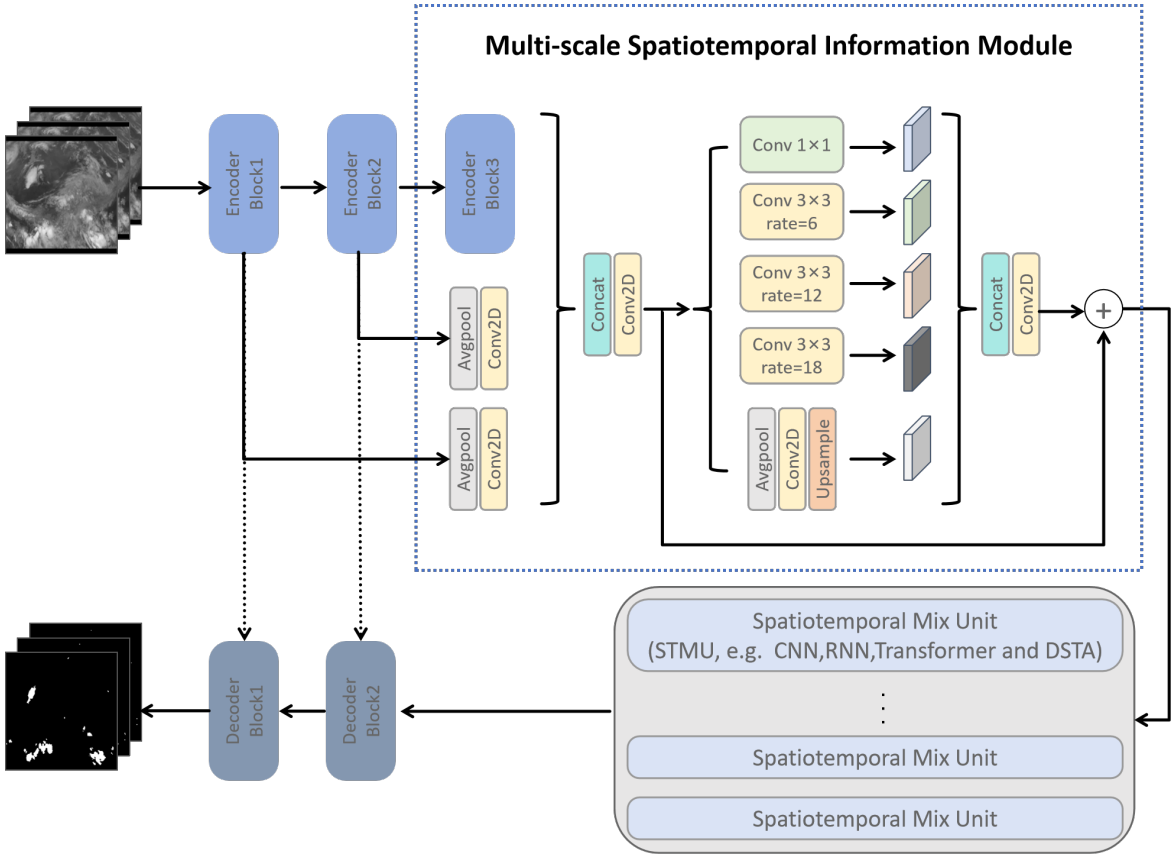



Fig. 2: Architecture of the proposed MCSDNet. A sequence of images is processed in parallel by a shared convolutional encoder. We introduce multi-scale spatiotemporal information by incorporating the feature maps from different encoder levels. At the lowest resolution, several Spatiotemporal Mix Units(STMUs) explore the spatial and temporal dependencies between the different frames. In the decoder, semantic information is transferred from the encoder to decoder to enhance the detection results further.

the Himawari-8 advanced Himawari-8 imager (AHI) data by combining machine learning models, area overlapping, and Kalman filter algorithms.

In recent year, deep learning has gained great breakthrough in computer vision tasks [28]. Some researchers attempt to utilize deep learning model to detect MCS region. CloudFU-Net [29] is a fine-grained segmentation method for ground-based cloud images which is based on encoder-decoder structure. Compared to threshold methods and machine learning methods, it possess better generalization performance and efficiently segment different cloud genera. Convection-Unet [2] has a simple architecture and utilizes the Unet architecture for pixel-level convection detection. FTransUNet [30] incorporates shallow-level and deep-level features in a multilevel manner, which apply multi-modal information on RSI semantic segmentation. However, these methods mentioned above are designed for single-frame MCS detection. They ignore the dramatic change and complex development of MCS. There is no special design to extract temporal evolution in the life cycle of MCS.

### B. Semantic Segmentation

The semantic segmentation methods are designed to classify each pixel of an image and assign a semantic tag to each pixel. Unet [31] is a u-shaped architecture network. It consists of a contracting path and an expansive path. The contracting path downsamples the images and extracts features from them. The expansive path upsamples the images and classifies each pixel. Deeplabv3+ [32] is based on encoder-decoder architecture. It utilizes atrous convolution and atrous spatial pyramid pooling(ASPP) to extract multi-scale semantic information. In Swin Transformer [33], a pyramid-like architecture is adopted to address the multi-scale characteristics of objects. It uses Transformer for dense predicting. However, the semantic segmentation models are not designed for remote-sensing images, and they are not robust enough for MCS detection. These models take a single frame of image sequence as input and ignore cross-frames consistent features. Due to the development of MCS exhibits strong temporal patterns, we need to consider both spatial and temporal information when detecting MCS regions.

### C. Video Understanding

Video Understanding task aims to analyze each frames in video and then classify the video or generate the future frames. Accurate video understanding yield substantial benefits across a wide range of practical applications, including

climate change analysis [34] [18], forecasting human motion [35], predicting traffic flow [36], and representation learning [37]. Inspired by the success of long short-term memory (LSTM) [38] in sequential modeling, ConvLSTM [18] leverages convolutional neural networks to model the LSTM architecture and represents a pioneering contribution to the field of video understanding. PredRNN [39] introduces a spatiotemporal LSTM unit which can effectively extract and retains spatial and temporal representations concurrently. Its subsequent iteration, PredRNN++ [40], enhances this by incorporating gradient highway units and casual LSTM, enabling adaptive capture of temporal dependencies. With the development of vision transformer architecture, [41] [42] [43] apply different space-time attention strategies to extract multi-scale spatiotemporal information and achieve great performance in video understanding task. In [44], Tan *et al.* introduce SimVP, a simple architecture which is completely built upon convolutional networks without recurrent architectures. Without introducing any extra tricks and strategies, SimVP can achieve state-of-the-art performance on various benchmark datasets. Inspired by the breakthrough in video understanding, MCSDNet utilizes both spatial and temporal information to detect MCS regions.

### D. Vision Transformer

Transformer is a simple network architecture based solely on attention mechanisms, dispensing with recurrence and convolutions entirely [45]. Transformer is first proposed in natural language processing(NLP) tasks and gains significant improvements in various downstream tasks, *e.g.*, machine translation, name entity recognition and large language model. Motivated by the remarkable performance achievements of Transformer in natural language processing (NLP), researchers have tried to explore the application of Transformer to computer vision tasks and started the investigation on ViTs. ViT proposes a pure transformer which is applied directly to sequence of image patches and performs very well on image classification tasks. Swin Transformer [33] designs a pyramid-like architecture and utilizes shifted windowing scheme to improve the efficiency of self-attention computation. Although ViTs have gained a great performance in natural image analysis, they are not well-suited for satellite image analysis.

## III. Methodology

### A. Preliminaries

We consider the MCS detection task as RSI sequence segmentation task. Given a RSI sequence $X \in R^{T\times C\times H\times W}$, with T the length of the sequence, C the number of channels, and H × W the spatial extent, we aim to predict the segmentation result $Y \in R^{T\times C\times H\times W}$ of these images.

The model with learnable parameters $\theta$ learns a mapping $F_\theta : X^{T\times C\times H\times W} \longmapsto Y^{T\times C\times H\times W}$ by exploring both spatial and temporal dependencies. In our case, the mapping $F_\theta$ is a fully convolutional network(FCN)[46] trained to minimize the difference between the predicted segmentation result and the ground-truth segmentation result. The optimal parameters $\theta^*$ are

$$\theta^* = arg\min_{\theta} L(F_\theta(X), Y), \tag{1}$$

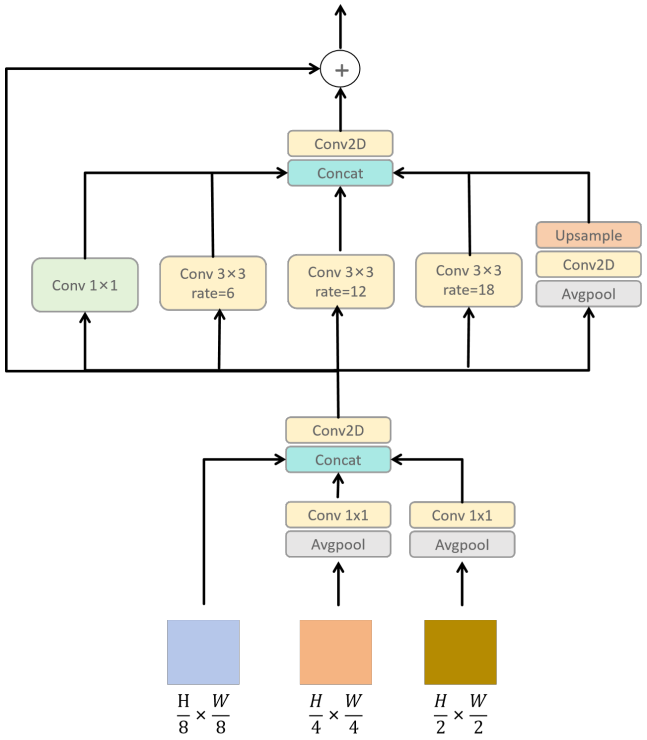



Fig. 3: Multi-scale spatiotemporal information module.

where L is a loss function that evaluates such differences. We adopt the FocalLoss as the loss function to calculate the loss between the predict result and the ground truth

$$L_{FocalLoss} = -\sum (1-p_t)^{\gamma} log(p_t), \tag{2}$$

where $p_t$ reflects the degree of closeness to the ground truth and $\gamma$ is adjustable factor. A larger $p_t$ indicates a closer proximity to ground truth.

### B. Overall Framework

Fig. 2 shows the architecture of the proposed MCSDNet. MCSDNet is based on an encoder–decoder architecture, which consists of three parts: an Encoder responsible for extracting spatial information, Spatiotemporal Mix Units (STMUs) responsible for modeling spatial and temporal information in image sequence, and a Decoder responsible for upsampling and reconstructing features from low-resolution feature maps.

In the encoder, MCSDNet adopts a RSI sequence $Z \in R^{T\times C\times H\times W}$ as input. Then, the model extracts the spatial features by several ConvNormReLU blocks(Conv2d+GroupNorm+ReLU). To reduce computational complexity, the different frames in sequence share the same multi-level spatial convolutional encoder. Each layer in encoder is formulated as

$$Z_i = \sigma(Norm2d(Conv2D(Z_{i-1}))), \tag{3}$$

where $\sigma$ is a non-linear activation function and Norm2d is a normalization layer. To maintain diversity between images at different time frames, we use Group Normalization with 4 groups instead of Batch Normalization in the encoder. Besides,

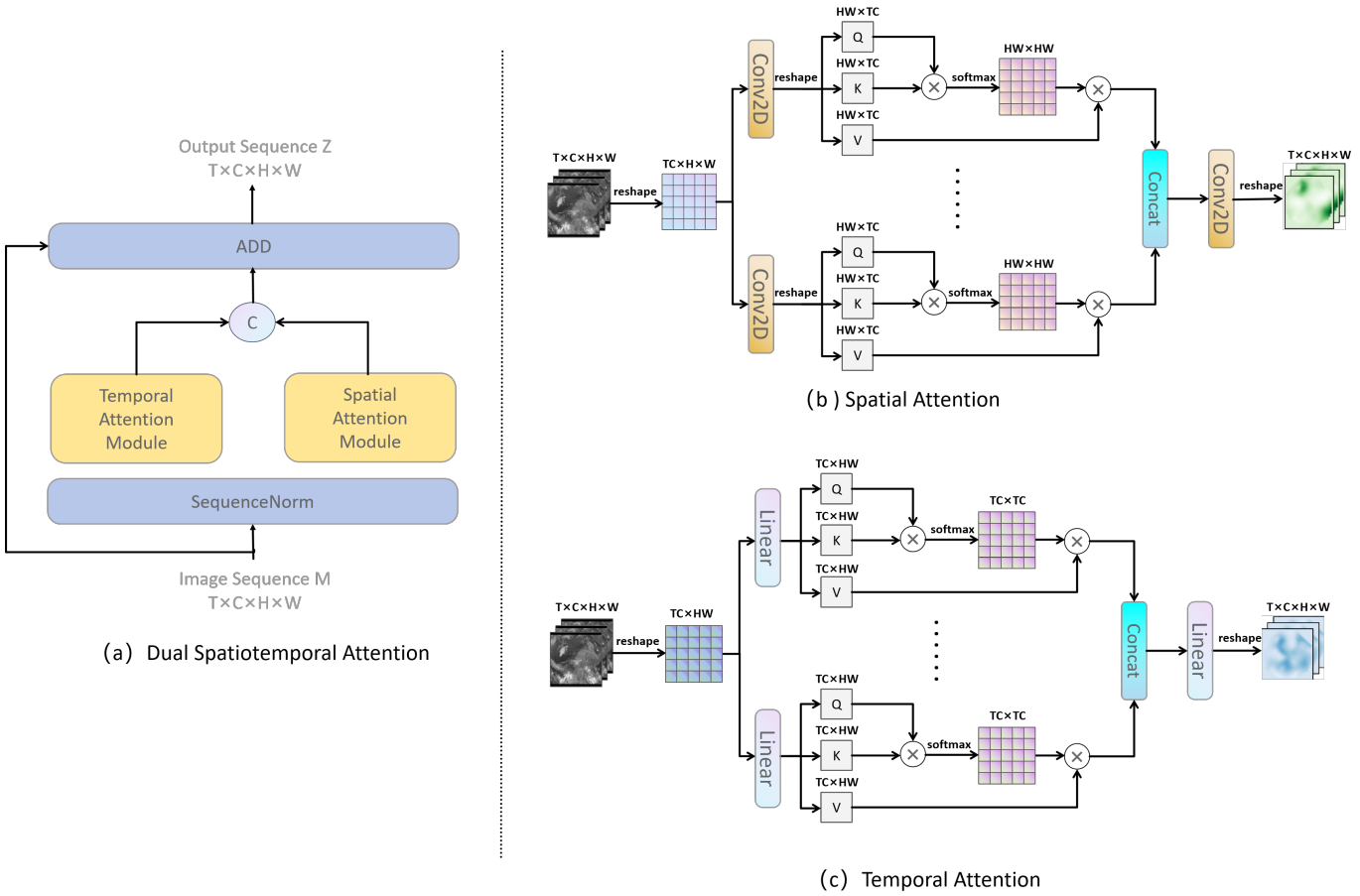



Fig. 4: The detail of proposed Dual Spatiotemporal Attention. (a) Architecture; (b) Spatial Attention Module; (c) Temporal Attention Module.

each layer in encoder begins with MaxPool with 2 scale factor to downsample the output of the front layer.

STMU is a scalable module, which can be replaced by other spatiotemporal modules. STMU takes the multi-scale spatiotemporal information from encoder as input and then explores the spatial and temporal dependencies between the different frames, which embeds the global context spatiotemporal information into image at different time. The development process of MCS exhibits strong temporal patterns. It can significantly improve the accuracy of MCS detection by introducing temporal information. In our implementation, we design a new module named Dual Spatiotemporal Attention(DSTA) as STMU. Different from the previous modules, DSTA can capture both intra-frame features and inter-frame correlations.

In the decoder, the model reconstructs the feature maps gradually. The lost spatial information in downsampling is addressed in the extremely lightweight decoder by using a skip connection from the encoder to the decoder. The detection accuracy is improved with a lower complexity increase [47]. Each layer in decoder is formulated as

$$Z_i = \sigma(Norm2d(Conv2D([e_i|unConv2D(Z_{i-1})]))), \quad (4)$$

where unConv2D is an upsmaple layer, $e_i$ is the output of the $i^{th}$ layer of the encoder, and [|] is the concatenation operation.

### C. Multi-scale Spatiotemporal Information Module

MCS is characterized by dense local distribution and sparse global distribution. We attempt to introduce multi-scale spatiotemporal information to improve the performance of MCS detection. As shown in Fig. 3, we design a multi-scale spatiotemporal information module to incorporate multi-level semantic of feature maps. In MCSDNet, we first apply global average pooling on high resolution feature maps from encoder and feed them into a $1\times1$ convolution to resize them into the same shape as the last feature map of encoder. With the global average pooling, we can reduce the loss of information during the process of changing the resolution of feature maps. Then, we concatenate the output from different encoder levels and pass them through a $1\times1$ convolution to generate a feature map with multi-scale spatiotemporal information.

The distribution of MCS is influenced by season and geographical location, making it more challenging to segment objects than that of natural images. The feature maps in different conditions have different optimal field-of-view. When detecting MCS regions, we desire more detailed spatiotemporal information. According to [48], we design a pyramid pooling module to extract multi-scale context information. With the pyramid pooling module, our model can adaptively adjust the field-of-view when keeping the resolution of feature maps, which captures MCS regions at multiple scales. The pyramid

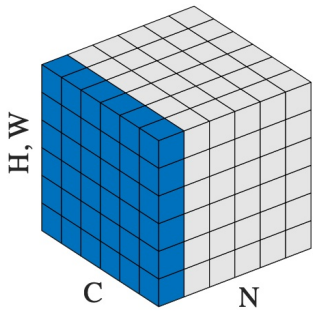



Fig. 5: Sequence Normalization which normalize the batch data along $T \times C \times H \times W$ dimension. T and C are concatenated together during normalization.

pooling module consists of several parallel atrous convolutions with different atrous rates and a global average pooling. We apply atrous convolutions to extract multi-scale information from RSIs which makes our model can capture MCS features from multiple field-of-views. Besides, the global average pooling incorporates global context information to the model and adopts the image level features. Then, we concatenate the results from all branches and feed them into a $1 \times 1$ convolutions to generate the final feature map.

### D. Spatiotemporal Mix Unit

MCSDNet applies Spatiotemporal Mix Unit(STMU) to capture the temporal evolution of MCS life cycle, which is important to detect MCS regions. Our proposed model has a simple architecture and is easy to expand. It can choose different STMU for sequence modeling, *e.g.*, CNN, RNN and Transformer. To effectively utilize the spatiotemporal features of MCS regions, we proposed a new STMU named Dual Spatiotemporal Attention(DSTA). Compared to the previous works [49] [20], DSTA simultaneously captures intra-frame features and inter-frame correlations. The previous STMU only take account into temporal evolution of MCS regions and ignores the texture features of convective cloud clusters which are important indicators to distinguish MCS regions and non MCS regions. In Section IV, we conduct an experiment to explore the performance of different STMU.

As shown in Fig. 4, DSTA consists of a temporal attention module(T-MSA) and a spatial attention module(S-MSA), which enable DSTA to extract both spatial features and temporal dependencies from RSI sequence. T-MSA performs self-attention operations along the temporal dimension, enabling it to model relationships across frames and capture the temporal evolution of MCS regions. Similarly, S-MSA performs self-attention operations along the spatial dimension, enabling it to capture long-range contextual semantic information and enhance the texture features of MCS regions. In addition, to maintain intra-sample diversity, we apply Sequence Normalization Fig. 5 to normalize the batch data, the normalization dimension is $T \times C \times H \times W$ , where T and C are concatenated together during normalization. Sequence Normalization is formulated as

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \varepsilon}} * \gamma + \beta, \tag{5}$$

where E[x] and Var[x] are mean and standard-deviation calculated over $T \times C \times H \times W$ dimension and $\gamma$ and $\beta$ are learnable parameters. By introducing DSTA, our model can not only consider the texture features of MCS regions at the current moment but also adopt the motion distribution of MCS regions in previous and subsequent moments.

DSTA adopts RSI sequence $X \in R^{T \times C \times H \times W}$ as input to explore the spatial and temporal dependencies between different frames and embeds the global context semantic into each frame. First, DSTA normalizes the RSI sequence along $T \times C \times H \times W$ to maintain the intra-sample diversity and accelerate the convergence speed of the model. Then, the parallel attention modules perform self-attention operations along TC dimension and HW dimension to capture intra-frame features and inter-frame correlations. Besides, we concatenate the outputs from two attention modules and feed them into a convolution layer to obtain better feature representation. To avoid performance degradation caused by gradients vanishing, the feature map after aggregation is reshaped to $T \times C \times H \times W$ and skip-connected to the input RSI sequence. The above process can be formulated as :

$$\begin{aligned} X^i = X^{i-1} + [&T - MSA(SequenceNorm(X^{i-1})) \\ &, S - MSA(SequenceNorm(X^{i-1}))], \end{aligned} \tag{6}$$

where $X^i$ is the output from DSTA, T-MSA is the temporal attention module, S-MSA is the spatial attention module and [ ] is the concatenate operation. With the relationship modeling capability provided by multi-head self attention, DSTA can efficiently embed global context semantic information into each frame in RSI sequence, leading to detect MCS regions more accurately.

## IV. Performance Evaluation

### A. Datasets and Settings

To evaluate the proposed MCSDNet, we conduct a comprehensive evaluation by comparing it with other RSI analysis, semantic segmentation and video understanding methods. These evaluations are performed on the newly introduced remote-sensing image datasets named MCSRSI. Besides, we design a sub-dataset of MCSRSI to explore the performance of the MCSDNet under extreme condition, *e.g.*, dense distribution of MCS regions. Then, we conduct ablation studies. The models are compared from both performance and efficiency aspects.

**Datasets** Due to the absence of publicly available MCS detection datasets derived from geostationary satellites, to evaluate the performance of MCSDNet, we create a dataset based on Fengyun-4A(FY-4A) for MCS detection task named MCSRSI. FY-4A is the first satellite of the FengYun-4 Geostationary meteorological satellite. The radiometric imaging channels of FY-4A increase from 5 on the Fengyun-2G(FY-2G) satellite to 14, covering visible light, shortwave

TABLE I: The distribution of images in MCSRSI.

| Season | Time | Train | Test |
|---|---|---|---|
| Spring | 2018-03 | 425 | 106 |
| | 2018-04 | 492 | 122 |
| | 2018-05 | 588 | 146 |
| Summer | 2018-06 | 516 | 129 |
| | 2018-07 | 520 | 130 |
| | 2018-08 | 627 | 156 |
| Autumn | 2018-09 | 525 | 131 |
| | 2018-10 | 536 | 133 |
| | 2018-11 | 632 | 157 |
| Winter | 2018-12 | 528 | 131 |
| | 2018-01 | 520 | 129 |
| | 2018-02 | 630 | 157 |
| Summary | ### | 6539 | 1627 |

infrared, mid-wave infrared, and long-wave infrared bands. It can capture the cloud map of the Earth every 15 min. MCSRSI is the first large-scale, publicly available dataset for MCS detection task based on visible channel. This dataset, as well as more information about its composition, are publicly available at https://github.com/250HandsomeLiang/MCSRSI.git.

MCSRSI is comprised of 8157 images in China region from FY-4A of shape $800 \times 1280$, ranging from January 2018 to December 2018 (see Table. I ). The time between acquisitions is 15 minutes. After calibrate the 12th channel of FY-4A's L1 data, we acquire visible channel image as model's input. With the help of domain experts, we utilize brightness temperature threshold and radar echo intensity to manually label MCS regions. MCS exhibits significant differences in brightness temperature in different regions. In the southern of China, MCS regions generally have lower brightness temperature threshold, ranging from 215K to 225K. However, in the northern of China, they tend to range from 220K to 235K, and even reach around 240K in some cases. In terms of radar echo intensity, convective cloud typically has radar echo intensity greater than 25dbz. Finally, we export ground-truth from origin image by setting the MCS pixel to 255 and the non MCS pixel to 0. As is common in remote sensing applications, MCSRSI is highly unbalanced, with the distribution of MCS regions in frame is less than 5% (see Fig. 1).

To get the RSI sequence for MCS detection, we utilize a 6-frame-wide sliding window with 30 min interval to organize image into a time-series sequence (see Fig. 6). In total, we obtain 7857 time-series sequences. Due to the distinct seasonal characteristics of MCS regions distribution and to prevent overfitting, we sample each month's data separately. First, we partition the data for each month into 5 equal groups, with one group used as the test set and the remaining four groups used as the train set. Then, we utilize the sliding window approach to generate time-series sequence for each group separately. The sequences that are not continuous in time are discarded. Through this approach, we obtain a train set with 6350 sequences and a test set with 1507 sequences. Both the train and test set contain time-series sequence from the entire year.

**Extreme Condition Evaluation** Mesoscale convective system, characterized by its abruptness and challenges in monitoring, poses significant threats to property and human lives. The unpredictability and rapid nature of convective events often leave authorities and individuals alike scrambling to respond, making it a particularly dangerous weather phenomenon. We hope that our model can achieve better performance in extreme conditions, especially in regions where MCS occurs frequently and in dense clusters. To this end, we have conducted an experiment to evaluate the performance of model across a diverse array of MCS distributions. The scope of this experiment is vast, encompassing a range of MCS occurrences spanning from the relatively rare 1% to the more common 5%.

**Experimental Setup** We implement the proposed method and other models in comparison with the Pytorch framework and conduct experiments on a single RTX 3090 GPU. We train MCSDNet, using Adam with a learning rate of 0.001, a batch size of 8 time-series sequences with 6 images and ReduceLROnPlateau as a scheduler to adjust the learning rate.

**Evaluation Metrics** We consider probability of detection (POD), False Alarm Ratio (FAR) and critical success index (CSI) as evalutaion metrics to compare MCSDNet with other methods. The evaluation metrics are formulated as

$$POD = \frac{TP}{TP + FN}, \tag{7}$$

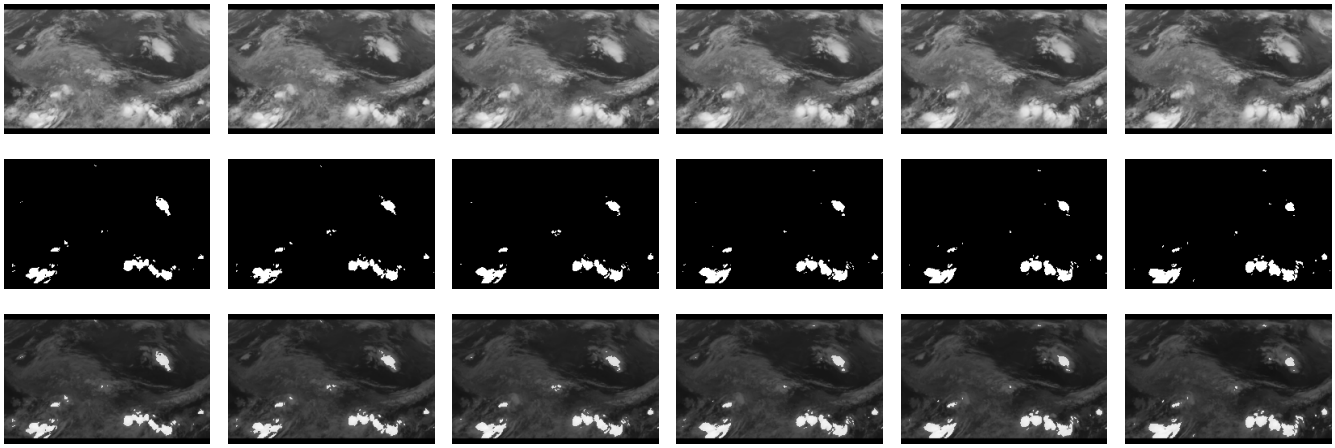
$$FAR = \frac{FP}{TP + FP}, \tag{8}$$

$$CSI = \frac{FP}{TP + FP + FN}, \tag{9}$$

where TP is true positive, FP is false positive, and FN is false negative. POD is used to measure the proportion of correctly detected MCS regions in the MCS set. FAR is used to measure the proportion of misclassified samples among those classified as MCS. CSI represents the IoU of positive samples, indicating the proportion of overlap between the positive regions in the ground truth and the prediction. Since POD and FAR are always inversely proportional, it is difficult to measure the quality of the model using these two metrics alone. Therefore, CSI is the primary metric used for performance evaluation, with POD and FAR as secondary metrics. A higher CSI and POD, and a lower FAR, indicate better model performance.

### *B. Performance Evaluation*

We evaluate our proposed method against strong baselines, including RSI analysis methods (CS-CNN [51], RS-NET [50]), semantic segmentation methods (Unet2D [31], DeepLabv3+ [32], Swin Transformer [33]), and video understanding (Unet3D, LSTMUnet, Video Swin [43], SimVP [44]).



(m) T=0 (n) T=30 (o) T=60 (p) T=90 (q) T=120 (r) T=150

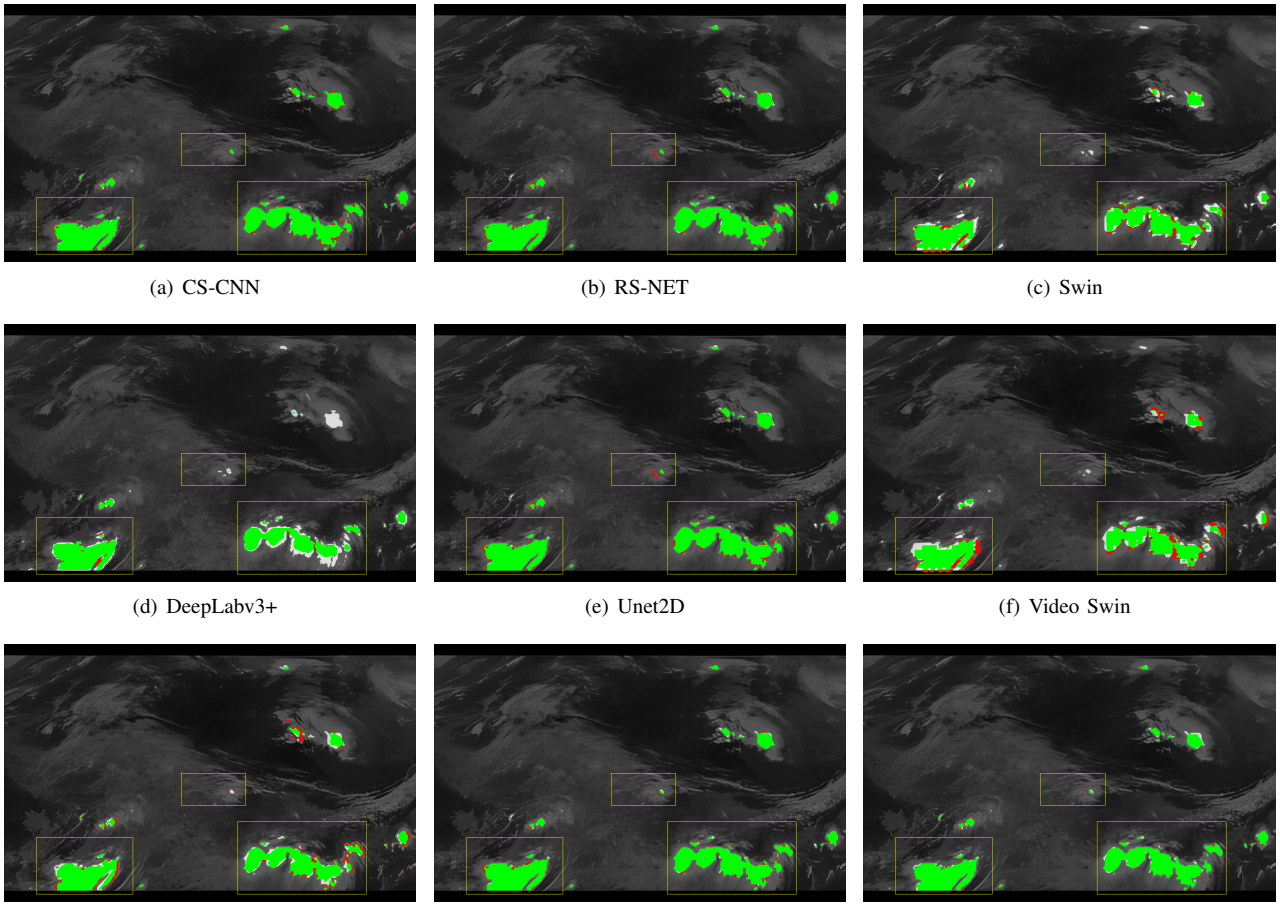
Fig. 6: A 6-frame-width time-series sequence in MCSRSI. From top to bottom, the order is as follows: input image, convective cloud label, and the distribution map of convective cloud.

TABLE II: Experiment results on MCSRSI. The best results are highlighted in bold. O:the model targets on single-frame. M:the model targets on multi-frames. R:remote sensing image analysis methods. S:semantic segmentation methods. V:video understanding methods.

| Method | Frames | Category | POD ↑ | FAR ↓ | CSI ↑ |
|---|---|---|---|---|---|
| RS-NET [50] | O | R | 0.85966 | 0.12059 | 0.77399 |
| CS-CNN [51] | O | R | 0.85829 | 0.10719 | 0.78452 |
| ConvectionUnet [2] | O | R | 0.86142 | 0.12234 | 0.77394 |
| Swin+UperNet [33] | O | S | 0.42113 | 0.25495 | 0.36962 |
| DeepLabv3+ [32] | O | S | 0.49900 | 0.21797 | 0.43248 |
| Unet2D [31] | O | S | 0.86545 | 0.13145 | 0.77201 |
| Video Swin [43] | M | V | 0.44182 | 0.28698 | 0.37855 |
| SimVP [52] | M | V | 0.77602 | 0.16065 | 0.73329 |
| LSTMUnet | M | V | 0.90183 | 0.16943 | 0.76884 |
| Unet3D | M | V | 0.87220 | 0.08692 | 0.81172 |
| **MCSDNet** | **M** | **R** | **0.92537** | **0.06530** | **0.87162** |

**Mesoscale Convective System Detection on MCSRSI Dataset** Table II summarizes the results of baseline models and MCSDNet on MCSRSI dataset. The results show that our proposed MCSDNet achieves state-of-the-art performance on MCS detection task when comparing with other methods. To be fair, all models are without a pretrain process. We can see that RSI analysis methods outperform semantic segmentation methods. It is because RSI analysis models are specifically designed for remote sensing imagery. They are more effective at extracting features from MCS compared to general semantic segmentation models. In semantic segmentation models, Unet has better performance than DeepLabv3+ and Swin Transformer. Texture feature is significant for MCS detection. Unet uses skip connections to capture multi-scale spatial information, allowing it to extract texture information from the image better. On the contrary, Swin Transformer and DeepLabv3+ focus on capturing the semantic information of images. Therefore, although they perform well in natural image semantic segmentation, they are not suitable for remote sensing image. The video understanding methods have better performance than other methods. This demonstrates that introducing temporal information can effectively improve the performance of MCS detection. Different from the above models, MCSDNet captures multi-scale spatiotemporal information from different encoder levels and apply a pyramid pooling to adaptively adjust the field-of-view. Then, our model utilizes DSTA to extract intra-frame features and inter-frame correlations. This not only emphasizes the texture feature of image but also learns the variation trend in the development process of MCS regions. MCSDNet integrates the advantages of RSI analysis models and video understanding models. By combining spatiotemporal data, it effectively improves the performance of MCS detection.

Fig. 7 shows the visual comparison of MCSDNet and other methods on the example from MCSRSI. Black represents the



(a) CS-CNN (b) RS-NET (c) Swin

(d) DeepLabv3+ (e) Unet2D (f) Video Swin

(g) LSTMUnet (h) Unet3D (i) MCSDNet

Fig. 7: Visual comparisons of different methods in MCSRSI. The white area represents convective cloud and the black area is non convective cloud. Besides, the green area represents true positive detection, while the red area represents false negative detection. (a) Result of CS-CNN (b) Result of RS-NET (c) Result of Swin Transformer (d) Result of DeepLabv3+ (e) Result of Unet2D (f) Result of Video Swin (g) Result of LSTMUnet (h) Result of Unet3D (i) Result of MCSDNet

background region, white represents the MCS region, green represents the correctly detected region, and red represents the region classified as convective cloud when it is not. From the results, we can see that MCSDNet has better performance than other methods. It has more true positive and less false positive detection. The intra-frame features and inter-frame correlations enable the model to capture long-term relations and improve the performance of MCS detection.

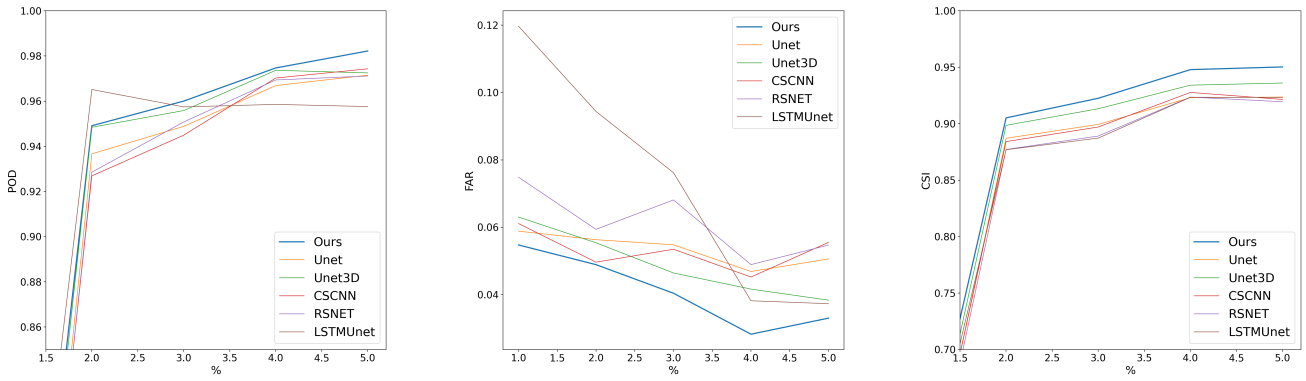
**Dense Distribution of Mesoscale Convective System** Fig. 8 presents a comprehensive comparison of the performance of our proposed model and the baseline model across various MCS distributions. All models are trained on MCSRSI with the same experiment setup. The results clearly illustrate the superiority of our model, especially in conditions characterized by frequent and dense convective weather events.

As the distribution of MCS regions increases, our model demonstrates a consistent and significant improvement in performance. Compared to the baseline models, our model gains better performance in extreme conditions, particularly when MCS regions occur frequently and densely. For example, when the distribution of MCS regions is 5%, our model achieves the best performance in all metrics. With a CSI of 0.95, our model demonstrates a high level of accuracy in identifying convective weather events. Additionally, a low FAR of 0.03 minimizes the occurrence of unnecessary warnings, while a high POD of 0.98 ensures that the majority of convective weather events are captured and not missed. This improvement can be attributed to the introduction of spatial and temporal information, which plays a crucial role in enhancing the model's ability to accurately detect MCS regions patterns and makes our model gain better performance in extreme conditions.

### *C. Ablation Studies*

We conduct ablation studies to evaluate the performance of MCSDNet and the proposed modules inside it. We train the models on the train set of MCSRSI and then test them on the test set. For performance evaluation, we compare the performance metrics such as POD, FAR and CSI.

**What's the role of multi-scale spatiotemporal information?** Due to MCS detection is a dense prediction task,




(a) Probability of Detection(POD) (b) False Alarm Ratio(FAR) (c) Critical Success Index(CSI)

Fig. 8: Performance curves of the models under different convective weather distributions. The blue line represents our proposed model.

TABLE III: The role of multi-scale spatiotemporal information

| Multi-scale | POD ↑ | FAR ↓ | CSI ↑ |
|---|---|---|---|
| | 0.88633 | 0.06031 | 0.84078 |
| ✓ | **0.92537** | **0.06530** | **0.87162** |

TABLE IV: The role of T-MSA and S-MSA in DSTA

| Temporal Attention | Spatial Attention | POD ↑ | FAR ↓ | CSI ↑ |
|---|---|---|---|---|
| | | 0.81650 | 0.32835 | 0.59543 |
| ✓ | | 0.84339 | 0.10359 | 0.78091 |
| | ✓ | 0.89667 | 0.09054 | 0.83123 |
| ✓ | ✓ | **0.92537** | **0.06530** | **0.87162** |

TABLE V: Performance of different STMU.

| STMU | POD ↑ | FAR ↓ | CSI ↑ |
|---|---|---|---|
| None | 0.81650 | 0.32835 | 0.59543 |
| Conv3D | 0.78002 | 0.28656 | 0.61255 |
| ConvLSTM | 0.95834 | 0.09352 | 0.87333 |
| Vision Transformer | 0.89667 | 0.09054 | 0.83123 |
| DSTA | 0.92537 | 0.06530 | 0.87162 |

we desire more detail spatiotemporal information to segment objects accurately. In MCSDNet, we concatenate the feature maps from different encoder levels and then apply a pyramid pooling to capture multi-scale spatiotemporal information. As shown in Table. III, the model with multi-scale spatiotemporal information gains better performance than that with single-scale information, with a 4% improvement in POD, a 3% decrease in FAR and a 3% enhancement in CSI. It shows that introducing multi-scale information can significantly improve the accuracy of MCS detection.

**Can we use other spatiotemporal modules as STMU?** In MCSDNet, we utilize STMU to capture the long-range temporal evolution of MCS regions from RSI sequence. With the rapid development of spatiotemporal modules, researchers maybe confused to choose which one as STMU [52]. In Table. V, we explore the performance of MCSDNet with different STMU, *e.g.*, CNN, RNN, Transformer and DSTA. From the results, we can see that the models with STMU detect MCS regions more accurately. Our proposed model exhibits remarkable scalability, facilitating the seamless integration of various spatiotemporal modules as STMU. Besides, MCSDNet maintains the original performance standards of the spatiotemporal modules without any compromise.

We design a new spatiotemporal modules named DSTA as STMU. Different from the previous modules, our proposed DSTA captures both intra-frame features and inter-frame correlations, and achieves the best performance among all STMUs.

**What's the role of attention modules in DSTA?** In our implementation, MCSDNet employs DSTA as STMU to capture spatiotemporal information of MCS regions. DSTA consists of a temporal attention module(T-MSA) and a spatial attention module(S-MSA). As shown in Table IV, when comparing with the baseline model, the models with attention modules improve the performance remarkably. The model with only T-MSA outperforms the baseline by 18.55% in CSI and 22.48% in FAR. T-MSA aims to capture the temporal evolution of RSI sequence and locate the MCS regions with cross-frames temporal consistency, which improves the FAR of MCS detection task. Meanwhile, employing S-MSA individually improves the detection accuracy from 81.65% to 89.67% in POD and 59.54% to 86.86% in CSI. S-MSA can emphasizes the texture features and distinguish MCS regions from none MCS regions better. When we integrate

TABLE VI: The optimal time interval in RSI sequence.

| Interval | POD ↑ | FAR ↓ | CSI ↑ |
|---|---|---|---|
| 15 min | 0.85548 | 0.09623 | 0.79255 |
| 30min | **0.92537** | **0.06530** | **0.87162** |
| 60min | 0.86778 | 0.09234 | 0.80540 |
| 120min | 0.85667 | 0.09510 | 0.79666 |

TABLE VII: Performance of network with combination of the individual module.

| Backbone | STMU | Decoder | POD ↑ | FAR ↓ | CSI ↑ |
|---|---|---|---|---|---|
| ✓ | ✓ | | 0.08112 | 0.37670 | 0.07539 |
| ✓ | | ✓ | 0.81650 | 0.32835 | 0.59543 |
| ✓ | ✓ | ✓ | **0.92537** | **0.06530** | **0.87162** |

the two attention modules together, the model with both T-MSA and S-MSA achieves the best performance in all evaluation metrics, with 92.54% in POD, 6.53% in FAR and 87.16% in CSI. From the results, we can see that introducing spatiotemporal information is significant to improve the performance of MCS detection task.

**What is the optimal time interval for image sequence?** It is important for MCS detection to introduce spatiotemporal information. We should choose appropriate time interval when designing RSI sequence. If time interval is too small, the changes in MCS regions are not significant and we can not effectively capture temporal evolution. Conversely, when the time interval is too large, there is a lack of correlations between consecutive frames. We conduct ablation study in four time interval: 15min, 30min, 60min and 120min. As shown in Table VI, when the time interval is 30min, MCSDNet achieve the best performance with 92.54% in POD, 6.53% in FAR and 87.16% in CSI. From the results, we find that time interval in RSI sequence significantly affects the performance of MCSDNet. As the time interval increases, the performance of model first improves and then declines. This suggests that both too small and too large time interval are not conducive to capture intra-frame and inter-frame variation.

**What roles do the Encoder, STMU and Decoder play?** The proposed MCSDNet consists of encoder, STMU, and decoder. Although our model has a simple architecture, it can improve the performance of MCS detection remarkably. We are eager to know the modules in MCSDNet how to influence the performance of MCS detection. As shown in Table VII, we find that STMU effectively improves the performance of model for MCS detection. Without STMU, the performance of model drops by 10.89% in POD, 26.31% in FAR and 27.62%. It is important for MCS detection to introduce spatiotemporal information. In MCSDNet, the encoder downsamples the origin image and extracts the spatial and texture information of MCS, and then the decoder gradually restores the feature map from encoder to the size of input frame. Besides, with skip connection, the lost spatial information during downsampling is compensated to the input of each decoder layer. We see that the performance of model without decoder significantly deteriorates so that it can not complete the MCS detection task. STMU and Decoder are essential for model to perform well on MCS detection.

## V. Conclusion

In this paper, we have proposed a novel deep learning model MCSDNet for MCS detection. MCSDNet has a simple architecture and is easy to expand. Different from previous methods, MCSDNet utilizes the multi-scale spatiotemporal information for MCS detection for the first time, which is more suitable to extreme conditions, especially when convective weather occurs frequently. In MCSDNet, we use STMU to capture temporal evolution in life cycle of MCS. STMU is a scalable module, which can be replaced by other spatiotemporal modules. It means that the future works about spatiotemporal modules can be easily migrated to our model. Besides, we design a new spatiotemporal modules named DSTA, which can capture both intra-frame features and inter-frame correlations. Then, we create a new dataset MCSRSI, the first large-scale dataset for MCS detection based on visible channel image. Evaluated on MCSRSI, our proposed MCSDNet gains state-of-the-art performance over existing MCS detection, semantic segmentation and video understanding methods. We hope that the combination of our open-access dataset and promising results will encourage the future research for MCS detection task, whose economic and environmental stakes can not be understated.

## References


[1] T. Fiolleau and R. Roca, "An algorithm for the detection and tracking of tropical mesoscale convective systems using infrared images from geostationary satellite," *IEEE transactions on Geoscience and Remote Sensing*, vol. 51, no. 7, pp. 4302–4315, 2013.

[2] Y. Wang and B. Xiao, "Convection-unet: A deep convolutional neural network for convection detection based on the geo high-speed imager of fengyun-4b," in *2023 International Conference on Pattern Recognition, Machine Vision and Intelligent Algorithms (PRMVIA)*. IEEE, 2023, pp. 163–168.

[3] Y. Yang, C. Zhao, Y. Sun, Y. Chi, and H. Fan, "Convective cloud detection and tracking using the new-generation geostationary satellite over south china," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[4] Z. Zhu and C. E. Woodcock, "Object-based cloud and cloud shadow detection in landsat imagery," *Remote sensing of environment*, vol. 118, pp. 83–94, 2012.

[5] Y. Zuo, Z. Hu, S. Yuan, J. Zheng, X. Yin, and B. Li, "Identification of convective and stratiform clouds based on the improved dbscan clustering algorithm," *Advances in Atmospheric Sciences*, vol. 39, no. 12, pp. 2203–2212, 2022.

[6] B. Utsav, S. M. Deshpande, S. K. Das, and G. Pandithurai, "Statistical characteristics of convective clouds over the western ghats derived from weather radar observations," *Journal of Geophysical Research: Atmospheres*, vol. 122, no. 18, pp. 10–050, 2017.

[7] M. Le Goff, J.-Y. Tourneret, H. Wendt, M. Ortner, and M. Spigai, "Deep learning for cloud detection," in *8th International Conference of Pattern Recognition Systems (ICPRS 2017)*. IET, 2017, pp. 1–6.

[8] S. Li, W. Song, L. Fang, Y. Chen, P. Ghamisi, and J. A. Benediktsson, "Deep learning for hyperspectral image classification: An overview," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 9, pp. 6690–6709, 2019.

[9] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.

[10] Y. Huang, J. Peng, G. Zhang, W. Sun, N. Chen, and Q. Du, "Adversarial domain adaptation network with calibrated prototype and dynamic instance convolution for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024.

[11] E. Basaeed, H. Bhaskar, and M. Al-Mualla, "Supervised remote sensing image segmentation using boosted convolutional neural networks," *Knowledge-Based Systems*, vol. 99, pp. 19–27, 2016.

[12] S. Mohajerani and P. Saeedi, "Cloud-net: An end-to-end cloud detection algorithm for landsat 8 imagery," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 1029–1032.

[13] J. Yang, J. Guo, H. Yue, Z. Liu, H. Hu, and K. Li, "Cdnet: Cnn-based cloud detection for remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 8, pp. 6195–6211, 2019.

[14] C. Luo, Z. Zhang, H. Lin, B. Zhang, X. Li, T. Zhang, and Y. Ye, "A practical online incremental learning framework for precipitation nowcasting," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–14, 2024.

[15] Z. Zhao, X. Dong, Y. Wang, and C. Hu, "Advancing realistic precipitation nowcasting with a spatiotemporal transformer-based denoising diffusion model," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 62, pp. 1–15, 2024.

[16] X. He, Y. Zhou, J. Zhao, D. Zhang, R. Yao, and Y. Xue, "Swin transformer embedding unet for remote sensing image semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–15, 2022.

[17] L. Lv and L. Zhang, "Advancing data-efficient exploitation for semi-supervised remote sensing images semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024.

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

[19] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[20] C. Tan, Z. Gao, L. Wu, Y. Xu, J. Xia, S. Li, and S. Z. Li, "Temporal attention unit: Towards efficient spatiotemporal predictive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 770–18 782.

[21] R. S. Schumacher and R. H. Johnson, "Organization and environmental properties of extreme-rain-producing mesoscale convective systems," *Monthly weather review*, vol. 133, no. 4, pp. 961–976, 2005.

[22] A. M. Haberlie and W. S. Ashley, "A radar-based climatology of mesoscale convective systems in the united states," *Journal of Climate*, vol. 32, no. 5, pp. 1591–1606, 2019.

[23] D. Chen, J. Guo, D. Yao, Y. Lin, C. Zhao, M. Min, H. Xu, L. Liu, X. Huang, T. Chen *et al.*, "Mesoscale convective systems in the asian monsoon region from advanced himawari imager: Algorithms and preliminary results," *Journal of Geophysical Research: Atmospheres*, vol. 124, no. 4, pp. 2210–2234, 2019.

[24] X. Huang, C. Hu, X. Huang, Y. Chu, Y.-h. Tseng, G. J. Zhang, and Y. Lin, "A long-term tropical mesoscale convective systems dataset based on a novel objective automatic tracking algorithm," *Climate dynamics*, vol. 51, pp. 3145–3159, 2018.

[25] L. T. Machado, M. Desbois, and J.-P. Duvel, "Structural characteristics of deep convective systems over tropical africa and the atlantic ocean," *Monthly Weather Review*, vol. 120, no. 3, pp. 392–406, 1992.

[26] L. Machado, W. Rossow, R. Guedes, and A. Walker, "Life cycle variations of mesoscale convective systems over the americas," *Monthly Weather Review*, vol. 126, no. 6, pp. 1630–1654, 1998.

[27] M. Kim, J. Im, H. Park, S. Park, M.-I. Lee, and M.-H. Ahn, "Detection of tropical overshooting cloud tops using himawari-8 imagery," *Remote sensing*, vol. 9, no. 7, p. 685, 2017.

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

[29] C. Shi, Z. Su, K. Zhang, X. Xie, X. Zheng, Q. Lu, and J. Yang, "Cloudfu-net: A fine-grained segmentation method for ground-based cloud images based on an improved encoder-decoder structure," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2024.

[30] X. Ma, X. Zhang, M.-O. Pun, and M. Liu, "A multilevel multimodal fusion transformer for remote sensing semantic segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[32] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[33] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[34] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and f. Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[35] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-d-based human motion recognition with deep learning: A survey," *Computer vision and image understanding*, vol. 171, pp. 118–139, 2018.

[36] S. Fang, Q. Zhang, G. Meng, S. Xiang, and C. Pan, "Gstnet: Global spatial-temporal network for traffic flow prediction." in *IJCAI*, 2019, pp. 2286–2293.

[37] S. Jenni, G. Meishvili, and P. Favaro, "Video representation learning by recognizing temporal transformations," in *European Conference on Computer Vision*. Springer, 2020, pp. 425–442.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

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

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

[41] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

[42] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[43] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

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

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[46] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[47] C. Luo, S. Feng, Y. Quan, Y. Ye, X. Li, Y. Xu, B. Zhang, and Z. Chen, "Trcdnet: A transformer network for video cloud detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

[48] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[49] H. Wang, W. Wang, and J. Liu, "Temporal memory attention for video semantic segmentation," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 2254–2258.

[50] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote sensing of environment*, vol. 229, pp. 247–259, 2019.

[51] J. Drönner, N. Korfhage, S. Egli, M. Mühling, B. Thies, J. Bendix, B. Freisleben, and B. Seeger, "Fast cloud segmentation using convolutional neural networks," *Remote Sensing*, vol. 10, no. 11, p. 1782, 2018.

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

**Jiajun Liang** is currently pursuing the B.S. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. His current research interests include computer vision, spatiotemporal data mining, and machine learning.

**Baoquan Zhang** is currently an Assistant Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. He received the Ph.D. degree in Computer Science from Harbin Institute of Technology, Shenzhen, China, in 2023. His current research interests include meta learning, few-shot learning, and machine learning.

**Chuyao Luo** received the bachelor's degree in Internet-of-Things engineering from Dalian Maritime University, Dalian, China, in 2017, and the Ph.D. degree in computer science from the Harbin Institute of Technology, Shenzhen, China, in 2022. He is currently a Post-Doctoral Researcher with the Harbin Institute of Technology. His research interests include data mining, computer vision, time series data prediction, and precipitation nowcasting.

**Xutao Li** is currently an Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. He received the Ph.D. and Master degrees in Computer Science from Harbin Institute of Technology in 2013 and 2009, and the Bachelor from Lanzhou University of Technology in 2007. His research interests include data mining, machine learning, graph mining, and social network analysis, especially tensor-based learning, and mining algorithms.

**Yunming Ye** is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China. He received the PhD degree in Computer Science from Shanghai Jiao Tong University, Shanghai, China, in 2004. His research interests include data mining, text mining, and ensemble learning algorithms.

**Xukai FU** is the legal representative of Beijing Jinkai New Energy Environmental Technology Co., 2510442827@qq.com, Ltd. His research interests include data mining, text mining, and ensemble learning algorithms.