Segmentation of Convective Initiation Based on Spatio-Temporal Feature Joint Modeling

Runzhe Tao*, Rui Chen, Peibei Zheng, Zibo Hong
Nanjing University of Industry Technology, Nanjing, China

***Abstract*—As a key indicator of the occurrence of severe convection, convective initiation (CI) exhibits characteristics such as fragmentation, scale heterogeneity, and susceptibility to confusion with other cloud systems in single-temporal remote sensing imagery, posing significant challenges for accurate CI detection. Traditional threshold-based methods inadequately capture spatial representations and have limited generalization capabilities, while existing deep learning approaches fail to fully utilize the temporal correlation features of the same target cloud cluster, resulting in a high false alarm rate. To address these challenges, based on the physical laws of convective development, we propose a spatiotemporal feature fusion-based CI detection model, namely Ti-UHRNet. The model integrates three core designs: integrating digital elevation model geographic information at the input layer to quantify the topographic modulation on convective development and enhance the physical consistency of features; adopting U-HRNet embedded with attention-gated feature fusion as the backbone to extract multi-scale features efficiently, filter critical information dynamically, and retain high-resolution spatial details of convective clouds; and designing a multi-head self-attention-based TransTrack module with multi-temporal inputs to capture the dynamic evolution information of convective clouds within a 15-minute window, thereby distinguishing them from other cloud systems. Experimental results show that compared with several advanced 2D and 3D convolutional segmentation methods, Ti-UHRNet achieves the best performance in extracting the spatiotemporal features of rapidly developing convective cloud clusters. On the test set, it attains a probability of detection of 0.954, a false alarm rate of 0.082, and a critical success index of 0.879. Verified against ground-based radar echoes, the model enables effective early warning of severe convective weather at 15–30 minutes in advance.**

***Keywords*—Semantic segmentation; remote sensing imagery; convective initiation; spatiotemporal feature fusion**

I. Introduction

Severe convective weather induces urban waterlogging as well as secondary geological disasters such as flash floods, landslides, and debris flows, not only posing serious threats to the safety of life and property of urban residents but also creating significant hazards to the normal operation of road traffic [1], [2]. To effectively address these threats, precise monitoring and early warning of severe convective weather are particularly crucial.

Currently, convective weather monitoring primarily relies on radar data and satellite remote sensing observations. Ground-based weather radars determine the position, structure, and intensity of convective clouds by transmitting and receiving electromagnetic waves. When a Doppler weather radar initially detects a convective cloud with a reflectivity factor ≥35 dBz [3],[4], the cloud is identified as Convective Initiation (CI), and the occurrence of CI marks the beginning of severe convective weather. However, in regions with complex terrain, especially plateaus and mountainous areas, the distribution of meteorological stations and radar sites is sparse, leading to the existence of observation gaps. Additionally, the obstruction of radar beams by complex terrain makes radar echo data susceptible to interference. In contrast, although geostationary satellite imagery has lower spatial resolution compared to ground-based radar, it offers unique advantages in monitoring mesoscale to large-scale weather systems. With its extensive coverage, high temporal frequency, and continuous tracking capabilities, satellites can effectively capture key characteristics such as the movement path and intensity evolution of weather systems, providing significant support for the monitoring and early warning of severe convective weather [5].

Meteorological satellites such as the GOES-18, Himawari-8, and China's Fengyun (FY)-2 and FY-4 series have all utilized multichannel data for CI detection [6]. Traditional methods integrate infrared brightness temperatures and their spectral differences in the window region to monitor the thermal structure of the lower troposphere, thereby constructing threshold-based 'interest fields' that characterize the evolutionary features of convective clouds and selecting key indicators from these fields to achieve the detection of CI [7]. However, this method is highly dependent on specific satellite data, geographical regions, and climatic conditions. In complex terrain areas such as plateaus, differences in underlying surfaces and environmental fields lead to complex and variable convective mechanisms, making it difficult for linear models combined with fixed thresholds to achieve reliable CI detection, and the generalization ability of such models is unsatisfactory [8],[9],[10]. In recent years, neural networks have progressively replaced threshold-based methods due to their powerful feature representation capabilities. For example, Sun et al. developed the Rapidly Developed Convection Monitoring System (RDCMS) algorithm based on TV-L1 optical flow and a Backpropagation Adaptive Boosting (BP_Adaboost) neural network, utilizing data from FY-4A geostationary satellite [11]. This system achieves a lead time of 17–40 minutes, with a Probability of Detection (POD) of 80% and a False Alarm Rate (FAR) below 34%. Similarly, Zheng et al. employed Himawari-8 satellite imagery to construct a deep belief network (DBN)-based method for identifying severe convective clouds, effectively detecting cloud clusters at various developmental stages from initiation to dissipation. Compared to traditional threshold methods and support vector

*Corresponding author.

machines, this approach improves identification accuracy [12]. However, early deep networks often have a 'black box' feature learning process. Although they can enhance overall identification accuracy, they struggle to model and utilize the physical laws governing convective development (such as thermodynamic and dynamic processes), rendering the models vulnerable to high-confidence false alarms. Chen et al. proposed ResU-Deep, a deep learning framework designed to improve the trigger function of deep convection in tropical regions. The model effectively learns the complex, multi-scale interactions between environmental thermodynamic fields and convective initiation. Compared to traditional physics-based trigger functions, it better captures the spatial continuity of convective onset over tropical oceans and landmasses and significantly reduces the false alarm [13]. Fan et al. revealed the critical roles of water vapor and cloud-top height in the formation of CI. Based on the ResNet architecture, they established a CI forecasting model that effectively learns the intrinsic correlations between GOES-16 satellite observation features (such as cloud-top temperature and water vapor characteristics) and CI events. Compared to traditional black-box deep learning models for CI nowcasting, this model significantly enhances the reliability and interpretability of its predictions [14]. However, although models based on Convolutional Neural Networks (CNNs) excel at capturing spatial features, their core convolutional operation has the inherent limitation of a local receptive field. Given the fragmentation, irregularity, and dramatic scale variations exhibited by CI on satellite cloud images, such models struggle to establish effective global contextual dependencies. This leads to missed detections of small-scale CI regions with weak signals. Furthermore, most current mainstream deep learning methods simply stack multi-temporal data as input channels, failing to fully exploit the dynamic evolution patterns of convective cloud clusters within a 15-30 minute time window (such as cloud-top cooling rate and expansion trends). This limitation makes it difficult to effectively distinguish between convective cloud clusters that are genuinely developing and those that are morphologically similar but non-developing, thereby becoming a core bottleneck leading to high false alarm rates.

Deep learning methods, with their flexible architectures, are capable of directly extracting multiscale spatiotemporal features from multisource remote sensing data. By effectively modeling the complex nonlinear relationships between atmospheric motion and input variables, they circumvent the subjectivity inherent in manual feature selection and the complexity of parametric modeling required by conventional approaches [15]. However, three fundamental obstacles still exist in practical applications: 1) The scarcity and uneven spatial distribution of radar observation data, particularly in plateau and mountainous regions, make it difficult to construct sufficiently large sample datasets based on the precise definition of CI. 2) On geostationary satellite images with a spatial resolution of 4 km, CI exhibits a fragmented distribution, irregular geometries, and significant scale variations. Some small-scale CIs occupy only a few pixels, resulting in blurred boundaries and spectral confusion with other cloud systems in the imagery. These factors severely limit the completeness and boundary accuracy of CI detection. 3) Convection under complex terrain conditions is often triggered by mesoscale and microscale processes such as mountain-valley winds, slope heating, and topographic lifting. These systems are characterized by short life cycles, rapid development, and highly localized spatiotemporal features, presenting distinctive challenges for detection [16].

To address the aforementioned challenges, this study first leverages satellite product data and radar data to construct a sufficiently scaled training sample set for deep learning. Based on the radar data available in partial regions and the Final Precipitation Data from the Integrated Multi-satellitE Retrievals for Global Precipitation Measurement mission (IMERG), the FY-4A AGRI L2 Convective Initiation Product (CIX) released by the National Satellite Meteorological Center (NSMC) is optimized, thereby improving the annotation accuracy of CI labels. Subsequently, based on the physical mechanisms of convective formation, this study proposes a CI detection model named Time-UHRNet (Ti-UHRNet). Guided by the physical mechanisms of severe convection occurrence, the model is designed with a focus on multi-dimensional feature extraction and spatiotemporal feature fusion, and its core design details are as follows. At the input stage, geographic information is integrated with multichannel satellite observation data to capture the impact of topographic forcing effects on convection development and thus enhance physical consistency. At the feature extraction stage, the encoder adopts a parallel convolution structure to reduce the loss of detailed information, while extracting multiscale spatial features; the decoder embeds an attention gating mechanism into the upsampling path to dynamically weight the feature importance of different channels and spatial regions, which guides the model to focus on the key convection-related information and improves the recognition sensitivity to the core convective regions. At the feature fusion stage, the temporal evolution features of convection development are introduced from the temporal dimension to capture the dynamic variation law of convective cloud clusters within a 15-minute time window, thereby distinguishing CI from other cloud systems. Through the above designs, the model can effectively integrate static geographic information, dynamic evolutionary features of atmospheric motion, and spatiotemporal correlation information, ultimately achieving synergistic and accurate detection of CI.

The remainder of this study is organized as follows: Section II describes the datasets used and elaborates on the process of label optimization. Section III systematically presents the overall architecture of the Ti-UHRNet model and the design principles of its core modules. Section IV details the experimental setup and the results of ablation studies, along with corresponding analysis. Section V provides a discussion and compares the proposed method with existing approaches. Section VI summarizes this study and provides an outlook on future research directions.

II. Data Acquisition and Label Preparation

A. Fengyun 4A Satellite Data

Fengyun serial meteorological satellites have made significant contributions to fields such as meteorology, oceanography, agriculture, forestry, water conservancy,

aviation, navigation, and environmental protection [17]. This study uses the L1_SDR Data of FY4A Advanced Geostationary Radiation Imager (resolution of 4 km) downloaded from NSMC. The data contains 14 channels, of which six from the water vapor and longwave infrared bands, covering a wavelength range of 6.5 to 13.3 μm, are utilized. Data preprocessing includes radiometric scaling and equidistant latitude-longitude projection for the study area. In addition, we use digital elevation model (DEM) data as input parameters to integrate the influence of altitude on infrared channel observation data, as well as the modulation effect of terrain on convection occurrence and development.

B. CIX Product of Fengyun 4A Satellite Data

Due to the lack of sufficient radar data in the plateau region to support the generation of CI labels, this study utilizes the CIX as a base label for correction and refinement. The CIX product is developed based on the RDCMS algorithm and the data within the product mainly consists of a two-dimensional matrix of CI identification for each satellite scan point. An identification value of -1 indicates that the satellite detected CI occurrence at the corresponding scan point, -2 indicates a potential convective cloud cluster in its initial stage, and 0 indicates no convective cloud cluster. The visualization of the read data is presented in Fig. 1(a), where white pixels represent values of -1 and -2. The CIX product has been validated against radar data in regions with ample radar coverage, achieving a CI detection probability of 90% but with a false alarm rate of 35% [11].

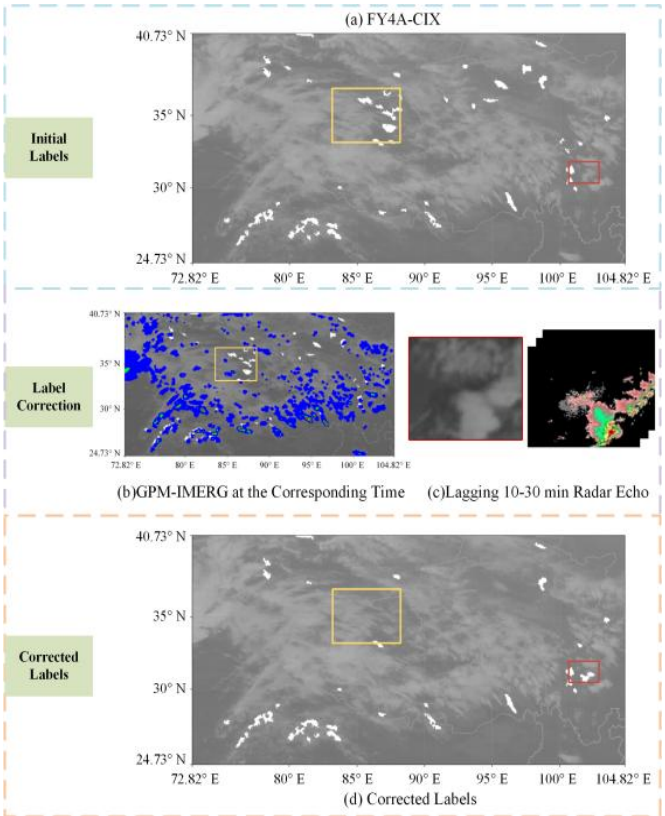

Fig. 1. The workflow of CI label preparation.

C. Global Precipitation Measurement and Weather Radar Echo

The target precipitation data used in this study is IMERG Final Run Data as shown in Fig. 1(b), which is a level-3 precipitation product of GPM. The data have a time resolution of 30 minutes and a spatial resolution of 0.1°. Final Run Data is based on direct statistical validation of the global ground survey station network [18]. During the verification process, radar and rain gauges from regions with different latitudes and climate characteristics around the world were combined, resulting in Final Run Data being the closest to actual precipitation, which provides reliable data support for various studies [19].

Weather Radar Echo data in Fig. 1(c) are sourced from the composite reflectivity image product of the national weather radar network provided by the National Meteorological Information Center, with a spatial resolution of 0.01° and a temporal resolution of 6 minutes.

D. Label Correction

To reduce the false alarm rate, this study calibrates the initial CIX product using the GPM-IMERG satellite precipitation product. The calibration is based on the principle that within mature convective systems, hydrometeors (e.g., raindrops, graupel, and ice crystals) eventually overcome updraft constraints and generate observable precipitation. Consequently, the presence of precipitation can serve as a reliable indicator of robust convection. Building on this premise, an elimination criterion is established: a CI candidate identified by the CIX product is flagged as a false alarm and subsequently removed if no precipitation is detected within an 8 km radius over the following hour. Although a small number of true CI events may be inadvertently eliminated, this method significantly reduces the false alarm rate. Validation was conducted in East China, where sufficient radar data are available, and the results show that eliminating spurious CI events based on actual precipitation reduces the false alarm rate to 7% while maintaining a hit rate of up to 90%.

Furthermore, to improve the detection probability of CI events in the labels, this study supplements the CIX product with radar data lagged by 10–30 minutes. This approach is based on the fact that satellites can capture signals from rapidly developing cumulus clouds approximately 10–30 minutes earlier than the appearance of 35 dBz echoes in radar observations [20]. However, due to the sparse distribution of radar stations over the plateau and mountainous areas, radar data are used as an auxiliary tool to enhance label quality. All data have been uniformly resampled to a spatial resolution of 4 km through interpolation-based upsampling and downsampling procedures. The final labels obtained after these two calibration steps are shown in Fig. 1(d). There are 1,818 samples covering the period from spring 2023 to spring 2024 in the dataset, of which 1,200 samples are randomly selected for training, 300 samples for validation and 318 samples for testing. Moreover, to ensure the validity of the experimental results, the training and testing data are maintained independent and all comparison results in the study are from the test set.

III. METHODOLOGY

The traditional CI detection primarily employs a multitemporal threshold method, which sets indicator thresholds by statistically analyzing the changes of the same cloud cluster over a period of time. For example, the channel brightness temperature difference (6.5μm channel minus 10.7μm channel) is used to determine the cloud-top height of tropospheric cloud clusters. A positive brightness temperature difference between water vapor and infrared channels indicates that the cloud top has reached or exceeded the tropopause, whereas a negative difference suggests a potential CI region. The channel brightness temperature difference (10.7 minus 12.5μm) helps distinguish cirrus clouds and deep convective clouds: a negative value corresponds to high, thin cirrus, while mature cumulus clouds in the upper troposphere exhibit positive values. The channel brightness temperature difference (10.7 minus 13.3μm) characterizes the developmental state of cumulus and cirrus clouds before convective initiation. Guided by these physical principles of convective development, this study employs a deep learning framework to simulate the CI process and formulates it as a multitemporal image segmentation task [21].

A. Establish the Model Guided by Physical Rules

This study designed a CI detection model named Ti-UHRNet based on spatiotemporal feature joint modeling, as illustrated in Fig. 2. First, DEM geographic information is fused with satellite multichannel data to effectively quantify the modulation effect of plateau topography (e.g., topographic lifting) on convective development, and high-quality multiscale spatial features of CI cloud clusters are extracted. Then, guided by traditional physical principles and taking advantage of the high temporal resolution of geostationary satellites, short-term temporal variation features during the formation of severe convection are mined from multiple infrared channels that are most highly correlated with CI. Finally, through spatiotemporal joint modeling, key dynamic information such as abrupt drops in cloud-top temperature and texture mutations is effectively captured, enabling accurate identification of CI in the complex plateau environment.

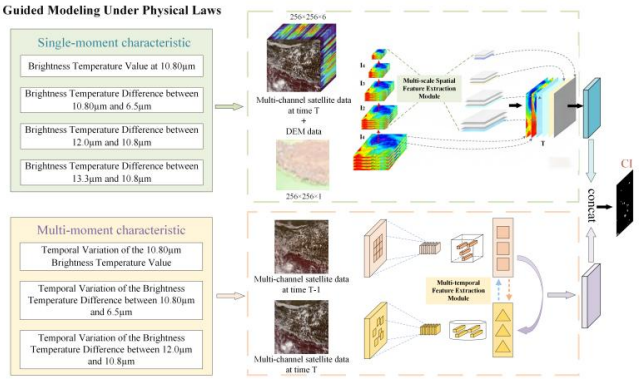

Fig. 2. CI detection modeling guided by physical principles.

For the proposed model, this study elaborates on the specific extraction methods of its spatial and temporal features in the subsequent sections. In the final spatiotemporal feature fusion stage, the model first integrates these two types of features via a feature map concatenation strategy, then performs feature enhancement by incorporating the Convolutional Block Attention Module (CBAM) that fuses channel and spatial attention mechanisms, and ultimately outputs the convection initiation detection results through this module.

B. U-HRNet for Spatial Domain Feature Extraction

Small clouds and fractus, which are typical manifestations of incipient CI, occupy only a few pixels in satellite images with a spatial resolution of 4 km × 4 km per pixel. This poses a critical challenge to the network's capability of extracting high-resolution spatial features, as the subtle brightness temperature variations and morphological characteristics of these small-scale CI targets are easily submerged by background noise and non-convective cloud clusters. During the feature extraction process, multilayer convolutional operations are widely adopted to capture the intrinsic characteristics of target objects; however, traditional multilayer convolutional networks tend to lose critical detailed information with the deepening of network layers due to repeated downsampling. This information loss is particularly detrimental to the spatial feature extraction of small-scale CI targets, where shallow-level spatial details (e.g., edge contours of incipient cumulus clouds) are equally essential as deep semantic features (e.g., cloud-top thermal characteristics) for accurate identification.

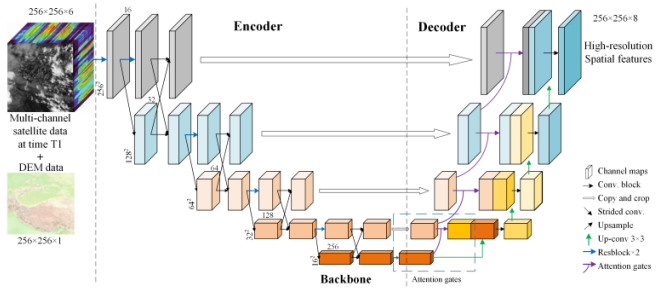

Fig. 3. U-HRNet for spatial domain feature extraction.

To address this dilemma, it is necessary to integrate low-resolution deep semantic feature maps with high-resolution shallow detail features through additional network layers. Meanwhile, high-resolution feature maps do not require excessive hierarchical processing, as each convolutional layer for high-resolution features incurs substantial computational costs, which would severely reduce the model's inference efficiency and hinder its practical deployment in meteorological operational services. Guided by this core demand—balancing fine-grained detail preservation, deep semantic representation, and computational efficiency. This study proposes an improved backbone network, namely U-High Resolution Network (U-HRNet), which is modified based on the original HRNet and tailored for CI detection. The structure of U-HRNet is illustrated in Fig. 3, and its effectiveness has been preliminarily verified in our recent study [22],[23].

Compared with the original HRNet, which is characterized by full-stage parallel convolution across multiresolution branches, U-HRNet introduces two targeted improvements to adapt to the unique characteristics of plateau CI detection, and

these improvements are quantitatively validated in subsequent ablation experiments (Section IV-B). 1) Optimization of network structure and computing configuration: Aiming at the redundancy and inefficiency caused by the full-stage parallel computation of multiresolution branches in the original HRNet, U-HRNet relaxes the parallel resolution constraint by retaining only the high-resolution feature stream in the initial stages and introducing additional processing stages solely after the deep low-resolution semantic feature map. It optimizes computing resource allocation by prioritizing low-resolution semantic streams (more capable of capturing deep semantic features) over high-resolution branches (mainly providing shallow spatial details), while achieving lightweight design through simplified parallel branches and optimized convolution kernel configuration. This design reduces redundant computation and model parameters while preserving high-resolution details, laying a foundation for the subsequent integration of the temporal feature extraction module (TransTrack) for spatiotemporal joint modeling. 2) To fully exploit the respective characteristics of multiscale feature maps extracted by the backbone network (high-level features contain rich semantic information but low localization accuracy, while low-level features achieve precise localization yet provide limited semantic information) and to overcome the drawback that traditional methods directly sum features pixel-wise without effective feature selection, which tends to introduce redundant information, U-HRNet introduces an Attention Gate module equipped with a soft attention mechanism in the feature fusion stage, as shown in Fig. 4.

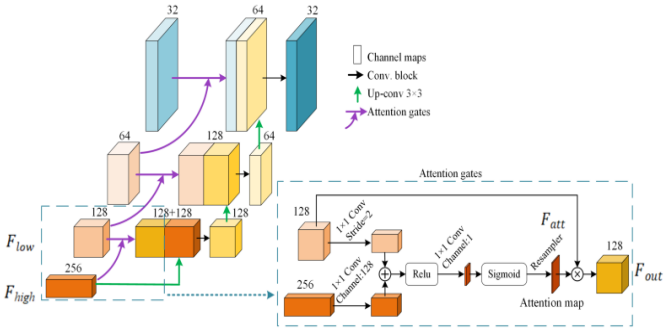

Fig. 4. Visualized structure of attention gates.

For the Attention Gate mechanism, this study utilizes low-level feature $F_{low} \in R^{C\times 2H\times 2W}$ position information to guide the resolution reconstruction of high-level $F_{high} \in R^{2C\times H\times W}$, achieving the fusion of different scale features [24]. Specifically, an Attention Gate is used to readjust the output characteristics on the high resolution before splicing the features on the high resolution with the corresponding fusion of the features on the low resolution. The Attention Gate generates a gated signal $F_{att} \in R^{1\times 2H\times 2W}$ that controls the importance of features at different spatial locations, allowing low-level spatial information and high-level detailed features to be fused layer by layer. The equations for calculation are Eq. (1) and Eq. (2):

$$F_{att} = Up\{Sigmoid\,[Conv_{1\times1}[Relu(Conv^{2}_{1\times1}(F_{low})+Conv_{1\times1}(F_{high}))]]\} \quad (1)$$

$$F_{out} = F_{att} \odot F_{low} \quad (2)$$

In the image upsampling process, this study utilizes transposed convolution to restore the scale of high-level features. The advantages of transposed convolution include adapting to the data through learnable parameters, such that the output not only enlarges the feature map but also reconstructs the input in the form of convolution. This is realized by a convolution kernel performing convolution operations on the feature map after it is expanded via zero-padding. These improvements enable U-HRNet to be more suitable for small-scale, weak-contrast CI cloud clusters over the plateau. It not only preserves shallow spatial details (critical for locating small CI targets) but also enhances deep semantic representation (essential for distinguishing CI from other clouds), thereby significantly improving the network's overall performance in balancing fine-grained detail perception and semantic feature extraction.

C. TransTrack for Temporal Domain Feature Extraction

The formation of convection is a temporally dynamic process evolving over time. To improve the reliability of CI detection results, it is necessary to fuse the high-quality spatial features extracted from cloud images with the information about atmospheric motion evolution derived from time-series data. For temporal feature extraction, this study designs an encoder-decoder structure, as illustrated in Fig. 5, to process global temporal information and constructs a cross-channel, cross-temporal target association mechanism. This approach fully leverages multi-channel observation data from adjacent frames at 15-minute intervals and by using the TransTrack module based on a multi-head attention mechanism to achieve multi-target cloud tracking [25], the development pattern of CI target cloud clusters can be extracted.

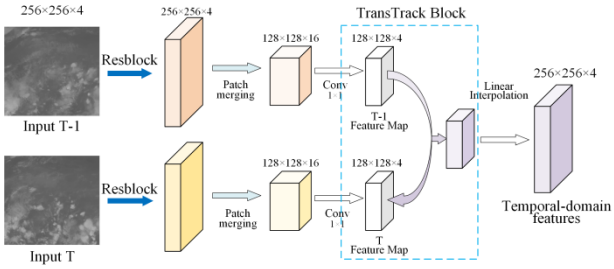

Fig. 5. TransTrack module for temporal domain feature extraction.

To effectively reduce computational complexity, and guided by physical mechanisms, this study selects four infrared channels (6.5 μm, 10.8 μm, 12.0 μm, and 13.3 μm) that are most relevant to the convective formation process as model inputs, with an input dimension of 256×256×4. The input data at time *t-1* is first subjected to patch merging and convolution operations, and the output is the feature map F_{t-1} with the size of 128×128×4. At this point, if F_{t-1} were flattened directly into a one-dimensional column vector and fed into the TransTrack encoding module, the computational complexity would be too great. The method of dividing into blocks is adopted, and the feature map F_{t-1} is split into 16 blocks ($L=16$), with each of size 32×32×4. Then, the network uses linear projection to obtain patch embedding (e_i) and encodes the spatial information of each block. The specific embedding position embedding (p_i) can be learned for block local position i,

forming the input sequence of *t-1* is $E_{t-1}=(e_1+p_1 \cdots e_L+p_L)$. In the same way, the input sequence E_t at time t can be obtained. Finally, E_t and E_{t-1} are fed into the encoder of the TransTrack module for temporal feature extraction.

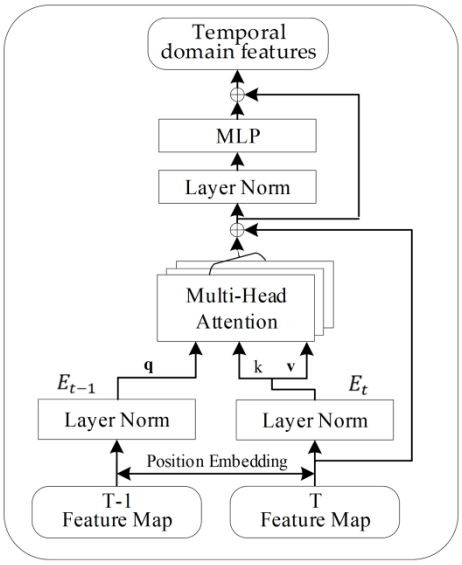

Fig. 6. Multi-head attention mechanism for multi-temporal data.

The TransTrack module leverages the self-attention mechanism of the Transformer architecture, which excels at long-range dependency modeling and target feature alignment. It first enhances and unifies the dimensions of the brightness-temperature feature maps from the four channels, then employs the Query-Key-Value mechanism to prepare the input for the multi-head attention component. The detailed structure of the multi-head attention layer is shown in Fig. 6, where the number of attention heads is set to 12. We use E_t and E_{t-1} to calculate the multi-head attention mechanism, where the Query matrix Q_{t-1} calculated by E_{t-1} is input into the calculation of the next frame sequence E_t, while the Key matrix as K_t and the Value matrix as V_t are calculated by E_t. The attention mechanism is defined as:

$$Attention(Q_{t-1},K_t,V_t)=softmax\left(\frac{Q_{t-1}K_t^T}{\sqrt{d_k}}\right)V_t \quad (3)$$

According to Eq. (3), the attention feature map can be obtained, where d_k is the dimension size of K_t, the scaling by $\sqrt{d_k}$ ensures gradient stability during training. Multiple attention heads are then concatenated to form the final multi-head self-attention output.

This design enables the Query matrix to retrieve relevant information between E_{t-1} and E_t. By computing cross-channel and cross-temporal correlation weights through the multi-head self-attention mechanism, the model can accurately identify and match the feature regions belonging to the same target cloud cluster in multi-channel data. This effectively mitigates the tracking discontinuity that plagues traditional methods when faced with dynamically evolving cloud morphologies and heterogeneous channel characteristics. As a result, the method uncovers temporally coherent feature maps that capture channel-wise variations of the same cloud cluster across consecutive observation times, thereby achieving efficient temporal information utilization.

IV. Experiments and Results

A. Implementation Details

The experiments adopt the Adam optimizer, which controls the weight distribution using an exponential decay rate with the coefficient 0.85, and controls the effect of the previous gradient's square by an exponential decay rate with the coefficient 0.999. The initial learning rate is set to 0.01 and the batchsize is set to 16. In the process of training, automatic learning rate decay is used. If the loss value of five successive epochs does not decrease, the learning rate will decrease by half, and the minimum value is 0.0001.

The loss function measures the quantification of the difference between the predicted value and the ground truth. The Cross Entropy Loss (CEL) function is usually used as a learning criterion to link to the optimization problem to solve and evaluate the model by minimizing the loss function. After counting all the samples, the ratio of background and CI pixels in the ratio of approximately 91:8. Since the background pixels are far more than CI pixels, the Dice Loss (DL) function is added to relieve the imbalance of the sample [26]. The total loss value is the weighted sum of the two loss function values. The calculation formulas are given as Eq. (4) to Eq. (6), where y_{label} is ground truth, $y_{predict}$ is the prediction of the model, and the weight coefficient α is set to 0.4.

$$CEL = y_{label} \times ln(y_{predict}) + (1 - y_{label}) \times ln(1 - y_{predict}) \quad (4)$$

$$DL = 1 - 2 \times \frac{|y_{label} \cap y_{predict}|}{|y_{label}| + |y_{predict}|} \quad (5)$$

$$Loss = \alpha \times CEL + (1 - \alpha) \times DL \quad (6)$$

To evaluate the performance of each semantic segmentation model, we adopt the Mean Intersection over Union (MIoU), and its corresponding equation is given as follows [see Eq. (7)]:

$$MIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{P_{ii}}{\sum_{j=0}^{k} P_{ij} + \sum_{j=0}^{k} P_{ji} - P_{ii}} \quad (7)$$

where, i denotes the true label, j is the predicted value, k denotes the number of categories (excluding the background). p_{ij} denotes the number of pixels that originally belonged to class i but are predicted as class j, p_{ii} denotes the number of pixels with correct predictions, p_{ij} and p_{ji} denote false positives and false negatives, respectively. In addition, based on the confusion matrix to quantify the performance of the CI detection model, TP denotes the number of pixels predicted as CI and actually being CI, FN denotes the number of pixels predicted as CI but actually being background, FP denotes the number of pixels predicted as background but actually being CI. The following three performance metrics are calculated: Probability of Detection (POD), Critical Success Index (CSI) and False Alarm Rate (FAR). The mathematical formulations are given as Eq. (8) to Eq. (10):

$$POD = \frac{(TP)}{(TP)+(FN)} \quad (8)$$

$$CSI = \frac{(TP)}{(TP)+(FP)+(FN)} \quad (9)$$

$$FAR = \frac{(FP)}{(TP)+(FP)} \quad (10)$$

Additionally, we employ the *Dice* coefficient, as shown in Eq. (11), which is a core metric for evaluating segmentation accuracy, especially in scenarios with class imbalance. It measures the overlap between predicted and ground-truth regions.

$$Dice = \frac{2\times TP}{2\times TP+FP+FN} \quad (11)$$

B. Ablation Experiment on the Spatial Feature Extraction Structure

This study selected U-Net++ as the baseline model, which exhibits outstanding performance in multiscale spatial feature extraction tasks. Relying on its densely connected skip connection structure, the model enables feature transmission and fusion, which not only alleviates the vanishing gradient problem but also excavates multilevel feature information in images more efficiently. It is thus particularly suitable for the segmentation tasks of data sources with high feature complexity and abundant detailed information, such as multichannel satellite images [see Fig. 7(a)] [27]. In the experiment, U-Net++ achieved an MIoU of 89.3% and a Dice coefficient of 90.5% on the test set for CI detection, as shown in Table I. However, constrained by its encoder-downsampling and decoder-upsampling serial architecture, U-Net++ not only incurs irreversible loss of high-resolution spatial details but also suffers from insufficient global context modeling, which is prone to result in erroneous detections. Meanwhile, the model adopts the method of directly concatenating features of the same scale for feature fusion. This approach fails to establish an effective correlation between the deep semantic features from the encoder and the shallow features from the decoder, thus making it difficult to bridge the semantic gap between them. As shown in Fig. 7(c), U-Net++ yielded misdetections for some non-convective initiation cloud clusters within the blue-circled region marked in Fig. 7(b), and its segmentation accuracy for the contours of target cloud clusters failed to meet the desired level. The low-resolution deep semantic features that characterized the global attributes of convective systems were easily overshadowed by the high-resolution shallow features containing massive background textures, ultimately leading to overly smooth segmentation results.

TABLE I. SUMMARY OF THE ACCURACY AND ERROR IN THE TEST SET FROM DIFFERENT METHODS.

Model	Miou	Dice	POD	FAR	CSI	Parameters
U-Net++	0.893	0.905	0.923	0.119	0.821	42.3M
HRNet	0.907	0.916	0.936	0.106	0.843	48.6M
HRNet+ Attention Gate	0.917	0.922	0.940	0.094	0.855	50.1M
U-HRNet	0.920	0.929	0.945	0.091	0.863	33.9M
Ti-UHRNet	**0.941**	**0.947**	**0.954**	**0.082**	**0.879**	**52.7M**
Ti-UHRNet (without DEM)	0.936	0.942	0.949	0.088	0.870	52.4M

The HRNet architecture is constructed upon a core architecture characterized by full-stage parallel convolution across multi-resolution branches and repeated cross-resolution feature fusion [28]. A salient advantage of this design is its ability to preserve a continuous high-resolution feature stream throughout the network, facilitating the synergistic learning of fine-grained spatial details and global semantic representations. This attribute substantially enhances target localization precision and multi-scale feature encoding capability, rendering the model well-suited for detection tasks with high spatial accuracy requirements. Concurrently, the elimination of elaborate upsampling operations for detail recovery effectively mitigates the risk of erroneous detections [29]. Experimental results demonstrate that the HRNet achieved a MIoU of 90.7% and a Dice coefficient of 91.6%. Compared to the baseline U-Net++ model, HRNet not only improves detection accuracy but also effectively reduces false detections. Nevertheless, the homogeneous cross-resolution fusion strategy fails to assign task-adaptive importance weights to features of different scales, leading to suboptimal capture of small-scale targets (e.g., weak incipient convective cloud clusters). As illustrated in Fig. 7(d), the model suffered from partial missed detections when identifying small-scale CI cloud clusters, which directly bottlenecked the further improvement of POD. This limitation ultimately restricted the overall detection performance of HRNet in practical CI monitoring scenarios. However, the paradigm of multi-branch parallel convolution incurs excessive model parameters and high computational complexity (with the parameter count reaching 50.1M). This not only limits the inference efficiency but also hinders subsequent joint spatiotemporal feature modeling.

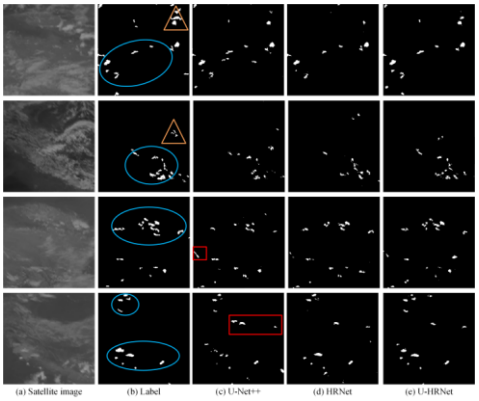

Fig. 7. CI detection results of different models: a) Satellite image at the corresponding time; b) Label as ground truth; c) Detection results of U-Net++; d) Detection results of HRNet; e) Detection results of U-HRNet.

The U-HRNet model proposed in this study demonstrated the best performance compared to other 2D convolutional networks, as shown in the blue-marked area of Fig. 7(e). This model innovatively integrates the core strengths of the two aforementioned networks: in the decoder, a lightweight parallel convolution structure is adopted, which not only enables efficient interaction between deep semantic features and shallow positional features but also effectively mitigates the redundant interference induced by cross-scale fusion; in the encoder, an attention module is employed to replace the traditional direct feature map concatenation operation, allowing for dynamic screening of key features, enhancement of feature

representation capability, and precise preservation of feature map resolution simultaneously. Experimental data showed that U-HRNet attained an MIoU of 92% and a Dice coefficient of 92.9%, representing a further performance improvement over HRNet. In terms of computational efficiency, U-HRNet had only 33.9M parameters, accounting for merely 69.7% of that of HRNet (48.6M). This effectively reduced the computational burden of the model and provided favorable conditions and feasibility for the joint modeling of temporal information in subsequent work.

A further analysis based on the three core evaluation metrics—POD, FAR and CSI—demonstrated that U-HRNet achieved a comprehensive leap in core CI detection performance while balancing detection accuracy and computational efficiency, thanks to its optimized feature fusion strategy and streamlined redundant parameters. Experimental results showed that U-HRNet attained a POD of 0.945, which was 2.2 percentage points higher than that of U-Net++ (0.923), 0.9 percentage points higher than that of HRNet (0.936), and 0.5 percentage points higher than that of the combination of HRNet and Attention Gate (0.940). This fully reflected its capability to accurately capture small-scale and weak incipient convective cloud clusters, effectively mitigating the missed detection issue. Its FAR was reduced to 0.091, which was 2.8 percentage points lower than that of U-Net++ (0.119), 1.5 percentage points lower than that of HRNet (0.106), and 0.3 percentage points lower than that of HRNet + Attention Gate (0.094), indicating that the model could effectively distinguish non-convective initiation cloud clusters from background noise through the feature screening of the attention module and the optimization of the lightweight structure, thus greatly reducing the misjudgment probability. Additionally, U-HRNet achieved a CSI of 0.863, representing an increase of 4.2 percentage points over U-Net++ (0.821), 2 percentage points over HRNet (0.843), and 0.8 percentage points over HRNet + Attention Gate (0.855). These results comprehensively reflected the advantages of U-HRNet in CI detection using single-temporal multichannel satellite data.

C. Ablation Experiment on the Temporal Feature Extraction Structure

CI features distinct meteorological characteristics of short duration and high abruptness, which render detection relying solely on single-temporal spatial features inherently limited. Specifically, for the CI targets in the orange triangular area indicated in Fig. 7(b), all models, including the baseline U-Net++, HRNet and U-HRNet, suffered from missed detections to varying degrees. In contrast, for the non-CI cloud clusters within the red rectangular box in Fig. 7(c), such 2D convolutional networks were prone to erroneous detection. This hindered the accurate discrimination of attribute differences between target and nontarget cloud clusters and thus failed to effectively mitigate the risk of misclassification. The core reason is that the brightness temperature characteristics of middle and low-level cloud systems such as thin cirrus clouds, stratocumulus clouds, and stratocumulus-cumulus mixed clouds are highly similar to those of CI clouds. Specifically, the infrared channel brightness temperature of thin cirrus clouds (with a cloud top height of 4 to 6 km) overlaps with that of CI clouds; the brightness temperature peak of stratocumulus clouds in areas with complex underlying surfaces overlaps with that of weak convective clouds; and the brightness temperature distribution of stratocumulus-cumulus mixed clouds is highly consistent with that of moderate-intensity convective clouds. It is difficult to achieve effective distinction relying only on the spatial features of a single frame image.

In terms of dynamic evolution rules, CI clouds show a short-term rapid development trend, which is specifically manifested in the rapid lifting of cloud top height, the sharp drop of brightness temperature, and the block-like expansion of cloud cluster morphology. In contrast, other similar middle and low-level cloud systems have relatively gentle temporal changes, with brightness temperature and cloud top height basically stable, and do not have explosive development characteristics or vertical development dynamic characteristics. Based on this, and guided by the physical laws of severe convective weather formation, this study introduced a temporal dimension feature fusion strategy into the 2D convolutional model framework. This approach aimed to capture the dynamic evolution information of the same cloud cluster within a 15-minute time window, thereby enabling refined monitoring and accurate identification of severe convective cloud cluster development processes. Meanwhile, considering the practical constraints of model computational efficiency and available hardware, and to seek an optimal balance between detection accuracy and inference speed, the input feature channels were specially selected to include the four key channels most relevant to the CI formation mechanism. This focus on core features was intended to avoid performance degradation due to redundant information and to further improve the model's practicality and deployment feasibility.

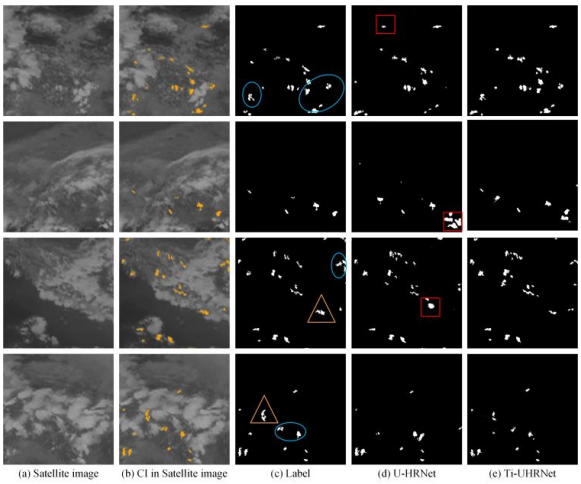

Fig. 8. The impact of temporal feature fusion on CI detection: a) Satellite image at the corresponding time; b) CI in satellite image; c) Label as ground truth; d) Detection results of U-HRNet; e) Detection results of Ti-UHRNet.

A comparison between the detection results in Fig. 8, where Fig. 8(d) and Fig. 8(e) reveals that the joint modeling of temporal features significantly enhances CI detection performance for the region marked in blue in Fig. 8(c). The extraction of multitemporal features is centered on the self-attention mechanism of the Transformer architecture, which excels at capturing long-range dependencies and aligning target

features across time. Through its multihead self-attention mechanism, the model computes correlation weights for cloud cluster features across different channels and temporal frames. This process enables accurate identification and matching of corresponding feature regions belonging to the same target cloud cluster within multitemporal data, thereby facilitating the construction of a cross-temporal feature interaction module. This module deeply fuses multichannel brightness temperature variation information with single-temporal spatial features, thereby providing comprehensive spatiotemporal and channel-wise three-dimensional feature support for the subsequent assessment of severe convection intensity and the prediction of its development trends. This significantly improves the accuracy and stability of target cloud cluster tracking.

On the test set, the Ti-UHRNet model achieved an MIoU of 94.1% and a Dice coefficient of 94.7%. Furthermore, for the false detection area within the red rectangle in Fig. 8(d), Ti-UHRNet substantially reduced false positives. It also successfully detected CI targets in the orange triangular area, which were missed when using only single-temporal spatial features, thereby further lowering the missed detection rate. In terms of key metrics, the Ti-UHRNet model achieved a POD of 0.954, reduced the FAR to 0.082, and increased the CSI to 0.879. Additionally, with a parameter count of 52.7M, Ti-UHRNet is only 4.1M larger than HRNet, ensuring improved detection accuracy without compromising efficiency. When DEM data were excluded from the input, detection performance declined slightly, verifying the effectiveness of incorporating DEM information for CI detection over complex terrain.

D. Validation of Detection Results in Areas with Sufficient Radar Data

The CI detection results are compared and verified with the observation data in areas with sufficient radar data to prove the generalization ability of the Ti-UHRNet model in practical applications. The Yunnan region and the Sichuan Basin were selected as the research areas, where frequent convective weather is driven by the significant influence of weather systems like the Southwest Vortex, combined with complex topographic conditions. Meanwhile, two typical cases in the spring and summer of 2025 were selected to elaborate on the model's detection performance in detail.

The first case occurred at 14:00 Beijing Time on July 10, 2025. The Ti-UHRNet model detected a CI signal at the central location of Chongqing. The red pixels in Fig. 9(a) represent the CI clouds identified by the model, with only a single target cloud appearing at this moment, featuring clear boundaries and a small area, indicating that the convective activity was in the initial triggering stage. The radar image at the corresponding time showed only blue to green echo responses in this area, with no signs of severe convective characteristics yet.

With the continuous development of the convective system, obvious changes occurred in the CI signals detected by the Ti-UHRNet model at 14:15 Beijing Time. Multiple CI target clouds were detected, and the coverage area of red pixels expanded significantly, indicating the intensification of the convective process. At the same time, orange and red pixel echo responses began to appear on the radar image, with the echo intensity increasing to 35–45 dBz and reaching more than 50 dBz in some areas, clearly indicating that severe convective weather had officially initiated. At this moment, the model detection results were in accurate agreement with the radar observation data. During the period from 14:30 to 14:45, the convective system further developed and intensified. The radar reflectivity image showed that within the CI area detected by the satellite at 14:15, the orange and red pixels increased explosively, the coverage area of high-intensity echoes (≥45 dBz) continued to expand, and the intensity of the echo center kept enhancing, with the echo intensity exceeding 55 dBz in some areas, indicating that the severe convective area was continuously expanding and the convective intensity was constantly upgrading. The above complete temporal process clearly demonstrates that the CI results detected by the Ti-UHRNet model using FY-4A multi-channel satellite data can capture the initial triggering signals of severe convective weather in advance, 15 minutes earlier than the initiation time of severe convection observed by radar and 30–45 minutes earlier than the peak time of severe convection, successfully realizing an accurate early forecast of severe convective weather and fully reflecting the practical value of the model in severe convection early warning.

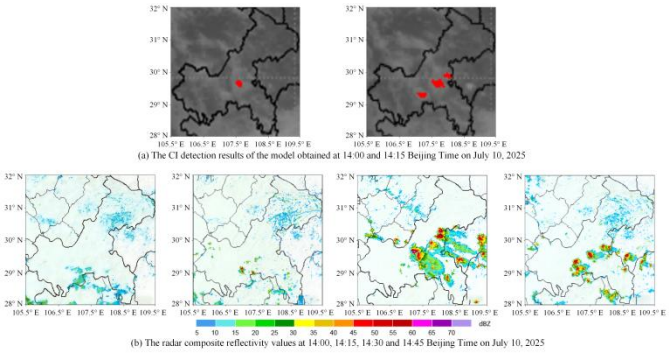

Fig. 9. Comparison of CI detection results in Chongqing area at 14:00 Beijing time on July 10, 2025 with radar echoes.

The second case occurred at 12:15 Beijing Time on May 21, 2025. The red signals in the model detection results were scattered in dot-like distribution, mainly concentrated in the northwestern part of Yunnan. It can be seen from the purple area in Fig. 10(a) that the area of each candidate CI target was small and the distribution was loose, indicating that the convective activity in this area was in a critical triggering state with strong uncertainty in convective development. At this moment, no echo response was observed on the radar reflectivity image corresponding to this area, that is, no signals related to convective activity were detected. It was not until 12:45, 30 minutes later, that a blue echo response appeared at the CI detection position marked by the purple circle on the radar echo image, with the echo intensity lower than 35 dBz. No continuous echo band was formed, and no severe convective weather developed. Therefore, the CI detection in the northwestern part of Yunnan this time was identified as a false alarm, mainly due to the instability and complexity of the CI process in this area, as well as certain differences between satellite cloud image observation and ground observation, leading to slight deviations in the model's identification of the critical state of convection.

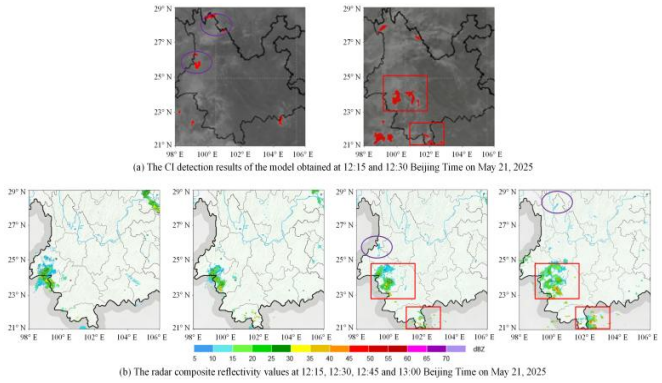

Fig. 10. Comparison of CI detection results in Yunnan area at 12:15 Beijing time on May 21, 2025 with radar Echoes.

Despite the existence of false alarms, the Ti-UHRNet model still showed outstanding performance in accurate detection in this case. At 12:30 Beijing Time, the Ti-UHRNet model detected a large-area, continuously distributed CI signal in the central part of Yunnan within the area marked by the red box on the satellite image. The red pixels were dense with clear boundaries, and the area of each CI target cloud was large, indicating that the severe convective activity in this area had entered a clear triggering stage and had the potential for continuous development. The radar reflectivity image at the corresponding time (12:30) showed an obvious increasing trend in the echo response in this area: the initial blue to light green echoes (intensity 15–25 dBz) gradually transformed into yellow echoes (intensity 25–35 dBz), and scattered orange pixel echoes (intensity 35–40 dBz) appeared, clearly indicating that the convective intensity began to gradually increase, which was highly consistent with the characteristics of the CI signals detected by the model. At 12:45 Beijing Time, the severe convective activity in this area further developed. More orange pixels appeared within the area marked by the red box on the radar reflectivity image, along with scattered red pixels (intensity≥45 dBz), and the coverage area of the CI signals detected by the model completely coincided with the severe convective area observed by radar. By 13:00, the radar echo response further enhanced, and a large area of continuously distributed red pixel echoes began to appear in the central and southern parts of Yunnan in the image, with the echo intensity generally reaching 45–55 dBz, indicating that the severe convective weather had entered the intense stage.

We can conclude that the CI signals detected by the Ti-UHRNet model in the two areas marked by the red boxes at 12:15 were completely accurate, and the model issued an early warning for severe convective weather in these two areas 15–30 minutes in advance compared with the initiation time of severe convection observed by radar (12:45). This effectively made up for the deficiency of slight false alarms and fully verified the detection reliability and early warning timeliness of the model in practical complex convective scenarios.

V. Discussion

Currently, many scholars conduct research on sequential image segmentation using 3D convolution [30]. However, 3D convolution exhibits relatively weak adaptability when capturing the features of short-term, rapidly evolving CI cloud clusters. Constrained by its convolutional kernel structure, it struggles to accurately capture the fine-grained changes of CI cloud clusters, such as morphological mutations (stretching, splitting, merging) and brightness temperature fluctuations, which are important indicators of CI initiation. This limitation restricts the application of 3D convolution in meteorological operational services. As can be seen from the indicators in Table II, compared to 2D convolution, although 3D convolution achieves a slight improvement in POD value, it does not reduce the FAR value. In comparison, the TransTrack module designed in this study has stronger pertinence in feature capture: it can more acutely capture the short-term variation laws of CI cloud clusters in the initiation stage, and effectively make up for the deficiency of 3D convolution in fine-grained dynamic feature capture. For the complex variation characteristics of cloud clusters during movement, the TransTrack module can real-timely track the brightness temperature value fluctuations, spatial distribution migration and intensity variation trends of cloud clusters in each channel, and extract the key dynamic features closely related to CI formation. These features provide reliable support for the accurate identification of CI, effectively detecting these fragmented small-scale CI cloud clusters.

TABLE II. Comparison of Accuracy and Computational Load between 3D Convolutional Networks and 2D Convolutional Networks.

Model	Miou	Dice	POD	FAR	CSI	Parameters
U-HRNet	0.920	0.929	0.945	0.091	0.863	33.9M
SegFormer	0.919	0.927	0.943	0.095	0.859	42.5M
3D U-HRNet	0.927	0.935	0.949	0.090	0.868	**87.2M**
3D UNETR	0.931	0.939	0.950	0.093	0.866	**76.8M**
Ti-UHRNet	**0.941**	**0.947**	**0.954**	**0.082**	**0.879**	**52.7M**

From the perspective of computational cost, 3D convolution has an inherent defect: its input is a stack of all infrared channels at two-time steps, and it is required to perform convolutional operations on both spatial and temporal dimensions simultaneously, which easily leads to an exponential growth in the number of model parameters and computational load, as shown in Table II. This defect greatly restricts the practical application of 3D convolution-based models by hardware computing power, making it difficult for such models to meet the real-time requirements of meteorological operational monitoring. In contrast, the Ti-UHRNet model is built on a 2D convolution architecture, which does not perform redundant convolutional operations in the temporal dimension and only fuses the short-term (15-minute) temporal features of CI cloud clusters in a targeted manner. Meanwhile, the model incorporates a channel screening strategy to eliminate redundant feature information, which significantly reduces the computational load on the premise of ensuring the effective capability to capture temporal features. This design concept enables the Ti-UHRNet model to achieve a favorable balance between monitoring accuracy and real-time performance, which is more in line with the actual demands of meteorological operational services. Additionally, this study selected UNETR, which is based on the Transformer

architecture, for comparative experiments to further highlight the advantages of the proposed Ti-UHRNet model.

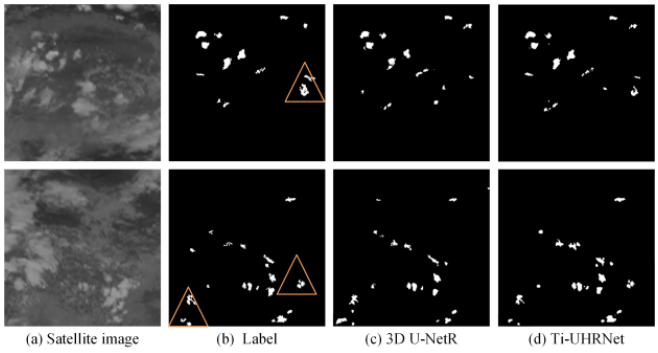

Fig. 11. Comparative detection results of 3D UNETR and Ti-UHRNet: a) Satellite image at the corresponding time; b) Label as ground truth; c) Detection results of 3D UNETR; d) Detection results of Ti-UHRNet.

The 3D UNETR model integrates the classic encoder-decoder structure of U-Net with the global feature modeling capability of Transformer, which has certain advantages in balancing local and global feature extraction: compared with pure Transformer models (such as ViT) [31], it retains local spatial details through the convolutional decoder; compared with pure convolutional U-Net models, its encoder enhances the ability of global feature perception (see Fig. 11). The comparative experiment showed that the UNETR (transformers for 3D Medical Image Segmentation) was always unable to effectively detect small-scale split cloud clusters within the orange triangle area in Fig. 11(b). This is because its attention mechanism in encoder is accustomed to capturing continuous entities with clear boundaries, whereas convective initiation cloud clusters in satellite imagery are characterized by tiny scales, fragmented distribution, and blurred boundaries. When the model performs global attention calculations on these sparse pixels, the weak target signals are easily diluted by background noise. As can be seen from Table II, Transformer-based models such as SegFormer and 3D UNETR, despite their inherent advantages in global feature modeling, encounter difficulties when processing small targets and object boundaries, with their CSI metrics being lower than those of their 2D and 3D U-HRNet counterparts. U-HRNet is specifically designed by us for small-scale cloud cluster segmentation. Therefore, it continues to serve as the backbone network for spatial feature extraction in Ti-UHRNet, while the Transformer architecture is added to capture local variation information, thereby achieving optimal detection results.

VI. Conclusion

Guided by the physical mechanisms of severe convection formation, this study proposes Ti-UHRNet, a novel deep learning framework designed to address the high-entropy problem of accurate CI detection on low-spatial-resolution satellite images. The model integrates three core designs: First, the integration of DEM geographic information into the model input effectively quantifies the modulation effect of terrain (e.g., topographic lifting, valley-mountain wind) on convection development, making the extracted features more physically consistent with the actual atmospheric process. Secondly, the U-HRNet backbone with an adaptive attention mechanism can effectively narrow the information bottleneck between high-level semantics and low-level spatial accuracy in remote sensing segmentation. This not only preserves the high-resolution spatial details of small-scale CI cloud clusters but also dynamically filters key convection-related features. Third, the TransTrack temporal feature extraction module based on multi-head self-attention captures the dynamic evolution characteristics of convective cloud clusters within a 15-minute time window, which makes up for the limitation of single-temporal spatial features that cannot distinguish CI clouds from similar middle-low level clouds (e.g., thin cirrus, stratocumulus). Through the integrated innovation of the model, we collectively tackle persistent challenges such as boundary ambiguity, spectral confusion, and the omission of small objects. Experimental validation demonstrates that this model significantly outperforms mainstream deep learning models such as U-Net++, HRNet, U-HRNet, SegFormer, 3D U-HRNet and 3D UNETR on core evaluation metrics. Furthermore, it achieves an advance warning lead time of 15–30 minutes for severe convection, providing a novel technical pathway and data support for severe convective weather warnings. The research results have important scientific significance for improving the level of plateau meteorological disaster prevention and mitigation, and also have important practical application value for the operational monitoring and early warning of severe convective weather in meteorological departments.

However, this study still has certain limitations that need to be further improved in subsequent research. First, the current model sets the temporal window of feature fusion to 15 minutes based on the FY-4A satellite observation frequency, and the adaptability of different temporal windows (e.g., 10 minutes, 20 minutes) to the rapid development of convection needs to be further explored. Second, the physical interpretation of the model's feature extraction process needs to be further deepened. Although the model design is guided by the physical mechanism of CI formation, the black-box characteristics of deep learning make it difficult to directly link the extracted high-dimensional features with specific atmospheric physical processes.

Acknowledgment

This study was funded by the Jiangsu Autonomous Driving Technology Innovation and Application Engineering Research Center Open Fund Project [grant number ZK24-06-05].

References

[1] Y. Wang, G. Gao, J. Zhai, Q. Liu, and L. Song, "Evolution characteristics of the rainstorm disaster chains in the Guangdong–Hong Kong–Macao Greater Bay Area, China," *Natural Hazards*, vol. 119, pp. 2011–2032, 2023.

[2] Y. Liu et al., "Flood impact on urban transport networks considering the flooding propagation," *Journal of Environmental Management*, vol. 395, p. 127972, 2005.

[3] Y. Zhang et al., "Skilful nowcasting of extreme precipitation with NowcastNet," *Nature*, vol. 619, pp. 526–532, 2023.

[4] R. D. Roberts and S. Rutledge, "Nowcasting storm initiation and growth using GOES-8 and WSR-88D data," *Weather and Forecasting*, vol. 18, pp. 562–584, 2003.

[5] M. Min et al., "Estimating summertime precipitation from Himawari-8 and global forecast system based on machine learning," *IEEE*

Transactions on Geoscience and Remote Sensing, vol. 57, pp. 2557–2570, 2018.

[6] D. Han, J. Lee, J. Im, S. Sim, S. Lee, and H. Han, "A novel framework of detecting convective initiation combining automated sampling, machine learning, and repeated model tuning from geostationary satellite data," *Remote Sensing*, vol. 11, p. 1454, 2019.

[7] J. R. Mecikalski, W. M. MacKenzie Jr, M. Koenig, and S. Muller, "Cloud-top properties of growing cumulus prior to convective initiation as measured by Meteosat Second Generation. Part I: Infrared fields," *Journal of Applied Meteorology and Climatology*, vol. 49, pp. 521–534, 2010.

[8] J. R. Walker, W. M. MacKenzie Jr, J. R. Mecikalski, and C. P. Jewett, "An enhanced geostationary satellite–based convective initiation algorithm for 0–2h nowcasting with object tracking," *Journal of Applied Meteorology and Climatology*, vol. 51, pp. 1931–1949, 2012.

[9] R. Fu et al., "Short circuit of water vapor and polluted air to the global stratosphere by convective transport over the Tibetan Plateau," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 5664–5669, 2006.

[10] R. J. Harris, J. R. Mecikalski, W. M. MacKenzie Jr, P. A. Durkee, and K. E. Nielsen, "The definition of GOES infrared lightning initiation interest fields," *Journal of Applied Meteorology and Climatology*, vol. 49, pp. 2527–2543, 2010.

[11] F. Sun, D. Qin, M. Min, B. Li, and F. Wang, "Convective initiation nowcasting over China from Fengyun-4A measurements based on TV-L 1 optical flow and BP_Adaboost neural network algorithms," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, pp. 4284–4296, 2019.

[12] Y. Zheng, X. Yang, and Z. Li, "Detection of severe convective cloud over sea surface from geostationary meteorological satellite images based on deep learning," *Journal of Remote Sensing*, vol. 24, pp. 97–106, 2020.

[13] M. Chen, H. Fu, T. Zhang, and L. Wang, "ResU-Deep: Improving the trigger function of deep convection in tropical regions with deep learning," *Journal of Advances in Modeling Earth Systems*, vol. 15, p. e2022MS003521, 2023.

[14] D. Fan, S. J. Greybush, E. E. Clothiaux, and D. J. Gagne, "Physically explainable deep learning for convective initiation nowcasting using goes-16 satellite observations," *Artificial Intelligence for Earth Systems*, vol. 3, p. e230098, 2024.

[15] C. J. Stubenrauch, G. Mandorli, and E. Lemaitre, "Convective organization and 3D structure of tropical cloud systems deduced from synergistic A-Train observations and machine learning," *Atmospheric Chemistry and Physics*, vol. 23, pp. 5867–5884, 2023.

[16] X. Xu, Y. Tang, Y. Wang, H. Zhang, R. Liu, and M. Zhou, "Triggering effects of large topography and boundary layer turbulence on convection over the Tibetan Plateau," *Atmospheric Chemistry and Physics*, vol. 23, pp. 3299–3309, 2023.

[17] J. Li, et al., "Quantitative applications of weather satellite data for nowcasting: Progress and challenges," *Journal of Meteorological Research*, vol. 38, pp. 399–413, 2024.

[18] A. Zhang, L. Xiao, C. Min, S. Chen, M. Kulie, C. Huang, and Z. Liang, "Evaluation of latest GPM-Era high-resolution satellite precipitation products during the May 2017 Guangdong extreme rainfall event," *Atmospheric Research*, vol. 216, pp. 76–85, 2019.

[19] X. R. Yang, S. K. Zeng, and Z. P. Lin, "Evaluation of monitoring ability of GPM satellite precipitation products for extreme precipitation over Sichuan Province," *Remote Sensing Technology and Applications*, vol. 38, pp. 1496–1508, 2023.

[20] J. R. Mecikalski and K. M. Bedka, "Forecasting convective initiation by monitoring the evolution of moving cumulus in daytime GOES imagery," *Monthly Weather Review*, vol. 134, pp. 49–78, 2006.

[21] Y. Lee, S. Min, J. Yoon, J. Ha, S. Jeong, S. Ryu, and M. H. Ahn, "Application of deep learning in cloud cover prediction using geostationary satellite images," *GIScience & Remote Sensing*, vol. 62, p. 2440506, 2025.

[22] R. Tao, Y. Zhang, L. Wang, Q. Liu, and J. Wang, "U-High resolution network (U-HRNet): cloud detection with high-resolution representations for geostationary satellite imagery," *International Journal of Remote Sensing*, vol. 42, pp. 3511–3533, 2021.

[23] R. Tao, Y. Zhang, J. Wang, P. Zheng, and L. Liu, "U-HRNet+: a real-time precipitation estimation method based on the fusion of temporal and spatial domain features," *International Journal of Remote Sensing*, vol. 45, pp. 4300–4323, 2024.

[24] Y. Gao, J. Guan, F. Zhang, X. Wang, and Z. Long, "Attention-unet-based near-real-time precipitation estimation from feng-yun-4A satellite imageries," *Remote Sensing*, vol. 14, p. 2925, 2022.

[25] T. Meinhardt, A. Kirillov, and L. Leal-Taixe, "TrackFormer: Multi-object tracking with transformers," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, Jun. 19–24*, 2022, pp. 8844–8854.

[26] R. Zhao et al., "Rethinking dice loss for medical image segmentation," *in Proceedings of the IEEE International Conference on Data Mining (ICDM), Sorrento, Italy, Nov. 17–20*, 2020, pp. 851–860.

[27] Y. Xu, B. Cao, and H. Lu, "Improved U-net++ semantic segmentation method for remote sensing images," *IEEE Access*, vol. 13, pp. 55877–55886, 2025.

[28] H. Wu, C. Liang, M. Liu, and Z. Wen, "Optimized HRNet for image semantic segmentation," *Expert Systems with Applications*, vol. 174, p. 114532, 2021.

[29] Y. Sun and W. Zheng, "HRNet-and PSPNet-based multiband semantic segmentation of remote sensing images," *Neural Computing and Applications*, vol. 35, pp. 8667–8675, 2023.

[30] G. Jin et al., "Spatio-temporal graph neural networks for predictive learning in urban computing: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, pp. 5388–5408, 2023.

[31] J. Wu, Z. Xu, X. Ai, and Z. Xu, "Three-Dimensional Reconstruction of Precession Warhead Based on U-NetR Network and Multi-Station HRRP Sequences," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 61, pp. 17828–17842, 2025.