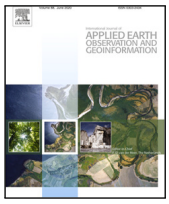
International Journal of Applied Earth Observation and Geoinformation 138 (2025) 104465

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag



VCDFormer: Investigating cloud detection approaches in sub-second-level satellite videos

Xianyu Jin [a], Jiang He [b,*], Yi Xiao [a], Ziyang Lihe [a], Jie Li [a], Qiangqiang Yuan [a,**]

[a] School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei, China
[b] Chair of Data Science in Earth Observation, Technical University of Munich, Munich, 80333, Germany

ARTICLE INFO

Keywords:
Satellite videos
Cloud detection
Deep learning
Spatial–temporal modeling

ABSTRACT

Satellite video, as an emerging data source for Earth observation, enables dynamic monitoring and has wide-ranging applications in diverse fields. Nevertheless, cloud occlusion hinders the ability of satellite video to provide uninterrupted monitoring of the Earth's surface. To mitigate the interference of clouds, cloud-free areas need to be selected before application, or an optimized solution like a cloud removal algorithm can be utilized to recover the occluded regions, both of which inherently demand the precise detection of clouds. However, no existing methods are capable of robust cloud detection in satellite videos. We propose the first sub-second-level satellite video cloud detection model VCDFormer to handle this problem. In VCDFormer, a spatial–temporal-enhanced transformer consisting of a local spatial–temporal reconfiguration block and a spatial-enhanced block is introduced to explore global spatial–temporal correspondence efficiently. Additionally, we construct WHU-VCD, the first sub-second-level synthetic dataset specifically designed to capture the more realistic motion characteristics of both thick and thin clouds in satellite videos. Compared to the state-of-the-art cloud detection methods, VCDFormer achieves an approximate 10%–15% improvement in the IoU metric and a 5%–8% increase in the F1-Score on the simulated test set. Experimental evaluations on Jilin-1 satellite videos, involving both synthetic and real-world scenarios, demonstrate that our proposed VCDFormer achieves superior performance in satellite video cloud detection tasks. The source code is available at https://github.com/XyJin99/VCDFormer.

1. Introduction

Satellite video has emerged as a promising data source for Earth observation, which can provide sub-second-level temporal information. Observing in a sit-and-stare manner, it has great potential in long-term analysis tasks like object tracking (Chen et al., 2024), dynamic monitoring (Xuan et al., 2019), and even military purposes. However, the quality of satellite videos is highly susceptible to cloud occlusion, which inevitably damages the long-term Earth's surface observation. To mitigate the impact of clouds, an intuitive solution is recognizing and selecting cloud-free areas before downstream application with the aid of corresponding cloud masks, yet will lead to a large amount of data lost and discarded. Another approach involves performing video cloud removal in satellite videos to recover a seamless record of the Earth's surface, where the precise location of clouds is also a necessary prerequisite (Jin et al., 2024). Overall, the accuracy of cloud detection plays a crucial role in improving the long-term monitoring capability of satellite video and beneficial to its further applications, thus a tailored video cloud detection (VCD) method is urgently required.

An intuitive approach is extending remote sensing (RS) cloud detection methods to satellite videos, in which significant advancements have been made over the past few years. Existing remote sensing cloud detection methods can be classified as traditional and deep learning-based algorithms. Tradition RS cloud detection methods mostly combine spectral (Zhu and Woodcock, 2012; An and Shi, 2015), textural (Deng et al., 2018), and geometric (Braaten et al., 2015) characteristics information for cloud detection in satellite images, yet lack generalization ability to different sensors and complex scenes. As for deep learning-based RS cloud detection methods, spatial and spectral information are explored to accurately classify cloud pixels (Francis et al., 2019; Li and Wang, 2024; Dong et al., 2024). However, almost all of them are based on mono-temporal satellite images. As satellite video contains only three basic channels RGB, its abundant temporal information and insufficient spectral information make the utilization of temporal information vital, especially in complex areas like thin clouds and high-reflection building roofs. Simply employing

* Corresponding author.
** Correspondence to: School of Geodesy and Geomatics, Wuhan University, Hubei, 430079, China.
E-mail addresses: jin_xy@whu.edu.cn (X. Jin), jiang.he@tum.de (J. He), qqyuan@sgg.whu.edu.cn (Q. Yuan).

https://doi.org/10.1016/j.jag.2025.104465
Received 7 January 2025; Received in revised form 18 February 2025; Accepted 4 March 2025
Available online 15 March 2025
1569-8432/© 2025 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Fig. 1. The difference between Fengyun4 images and satellite videos.

mono-temporal cloud detection methods on satellite videos will lead to inconsistent temporal results and severe false detection phenomena. As far as we know, there is recently one deep learning-based method TRCDNet (Luo et al., 2023) exploring temporal information to perform cloud detection. As shown in Fig. 1, TRCDNet is conducted on the Fengyun4 geostationary satellite images, in which the temporal sampling period is 15 min, and the cloud deformation across time is discontinuous and non-rigid. However, in sub-second-level satellite videos with much higher temporal resolution (0.04 s), cloud deformation is relatively minor, and cloud motion remains continuous across consecutive frames. This continuity makes the modeling of spatial–temporal relationships between frames crucial. Additionally, TRCDNet extracts spatial–spectral information independently in each frame with the aid of a near-infrared band, and then explores spatial–temporal correspondence in a single transformer layer, lacking full utilization of temporal information. In summary, due to the special characteristic of satellite videos, which can be denoted as dense temporal and deficient spectral information, no appropriate RS methods can be applied to VCD tasks currently.

Another solution is to draw inspiration from automatic video object segmentation (AVOS) methods in the field of computer vision (CV), which have been developed rapidly in recent years. As the mainstream architecture of AVOS, motion-based methods (Cho et al., 2023; Yuan et al., 2023) estimate optical flows between frames to integrate motion information. With the assistance of motion cues, salient objects can be accurately segmented from the background. Despite their promising performance in CV fields, the model performance would decrease when conducted on satellite videos due to the data gap between natural and satellite videos. As shown in Fig. 2, thin clouds in satellite videos make the modeling of optical flows between adjacent frames unstable. In that case, damaged flows might introduce interference and degrade the model. Besides, thin clouds and high-reflection areas make the imaging progress more complex, which brings extra difficulties in modeling temporal–spatial information in a motion-based manner. Other correlation-based methods (Lu et al., 2019; Karim et al., 2023) perform video segmentation by exploring spatial–temporal dependence implicitly without the aid of motion. However, unlike the natural objects which are salient and gathered, clouds in the satellite videos are distributed in a fragmented and random manner so that global correspondence and spatial–temporal position modeling are of even greater importance in video cloud detection tasks.

To address the current limitations in VCD, we propose the first sub-second-level architecture named Video Cloud Detection Transformer (VCDFormer). VCDFormer consists of four spatial–temporal-enhanced transformer (STET) layers to effectively capture inter-frame spatial–temporal dependencies at multiple scales. The STET leverages the transformer's capacity for long-term correspondence modeling to address challenges posed by fragmented cloud patterns commonly observed in satellite videos. To enhance the accuracy of dynamic cloud detection and reduce interference from complex imaging conditions, such as high-reflection areas, STET includes a simple yet effective component, the local spatial–temporal reconfiguration block (LSTRB). This module significantly enhances the detection of thin clouds and reduces false positives caused by high-reflective Earth's surfaces. Additionally, a spatial-enhanced block (SEB) is incorporated to model global spatial–temporal relationships within high-resolution satellite frames, enabling efficient and decoupled exploration of both local and global spatial–temporal information.

The contributions of this work are summarized as follows:

- We proposed the first sub-second-level satellite video cloud detection model VCDFormer, which leverages STETs to explore inter-frame spatial–temporal dependence across multiple scales.
- In STET layer, an effective local spatial–temporal reconfiguration block and a spatial-enhanced block are proposed to enhance the capability of capturing dynamic clouds meanwhile mitigating interference from complex imaging environments.
- To better characterize the motion properties of thick and thin clouds in real satellite videos, a synthetic dataset WHU-VCD based on the atmospheric scattering model with authentic cirrus cloud bands is proposed and released.
- As the first sub-second-level satellite video cloud detection model, VCDFormer demonstrates remarkable performance in both simulated and real-world experiments on Jilin-1 satellite videos, which performs the superior capability of detecting extremely thin clouds from complicated imaging conditions.

The rest of this paper is structured as follows. Section 2 introduces the related work to VCD tasks. Section 3 presents a detailed description of the proposed methodology. The creation process of the WHU-VCD dataset, along with the experimental results, is provided in Section 4, which also includes a few in-depth discussions. Finally, the conclusions and prospects of our work are summarized in Section 5.

2. Related work

Video cloud detection aims to detect cloud-covered pixels from corresponding cloudy areas. Given the lack of existing methods for sub-second-level satellite video cloud detection, we gain sights from deep learning-based RS cloud detection techniques and CV video segmentation algorithms.

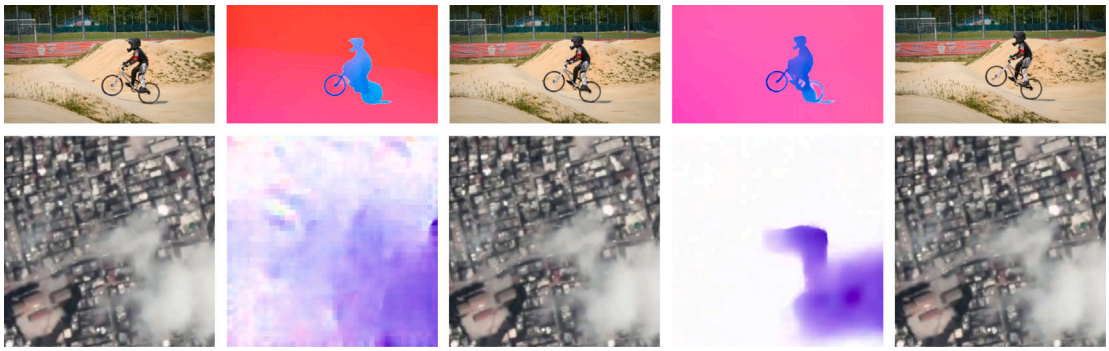


Fig. 2. The difference between natural and satellite videos, and the optical flow corruption in adjacent satellite video frames.

2.1. RS cloud detection

Various cloud detection methods have been developed in recent years, which can be broadly classified into traditional and learning-based approaches. Traditional methods mostly seek to design hand-crafted rules to achieve detection, relying on the physical characteristics of clouds and cloud shadows, such as spectral, textural, and geometric features or a combination of them. For instance, Irish et al. (2006) proposed the automatic cloud cover assessment (ACCA) by employing multiple determination rules based on spectral information in Landset-7 ETM+ images. The function of mask (Fmask) (Zhu and Woodcock, 2012) introduced the Landsat Top of Atmosphere (TOA) reflectance and brightness temperature (BT) as input to produce a probability mask for clouds, which is effective in Landsat 4–8 images. In Zhang et al. (2002), a haze-optimized transformation (HOT) was designed to exploit the visible-band space and quantify spatial variations in a 2-D spectral space. Others either employed geometric features like DEM (Braaten et al., 2015), or reference reflectance data as auxiliary data (Sun et al., 2016) for aid. Apart from them, a few methods (Zhu and Woodcock, 2014; Qiu et al., 2020) utilized temporal differences to perform change-based cloud detection, which takes cloud-free images as references. As manually designed rules are widely employed in traditional methods, the insufficiency of generalization ability makes them inflexible to complex scenes.

Benefiting from the rapid development of machine learning and deep learning, massive learning-based cloud detection methods have been widely investigated and improved. Some methods based on classical Bayesian (Hollstein et al., 2016), random forest (Ghasemian and Akhoondzadeh, 2018), and support vector machine (SVM) (Yuan and Hu, 2015) have shown promising performance in cloud detection, yet still lack the capability of exploring deep features. Leveraging the representational strengths of deep neural networks (He et al., 2023a,b), models employing CNN (Chai et al., 2019) and RNN (Mateo-García et al., 2019) are proposed for cloud detection. For example, Mohajerani and Saeedi (2019) designed a fully convolutional network (FCN) to detect clouds in Landset-8 images. Yang et al. (2019), Guo et al. (2020) proposed a series of CDNets to explore the spatial–spectral information to extract cloud masks efficiently. A deformable convolutional network named DCNet was introduced in Liu et al. (2021a) to handle the geometric transformations of clouds. Focusing on the boundary details of clouds, a modified segmentation method named Boundary-Nets is developed in Wu et al. (2022). While mainstream algorithms commit to designing elaborate modules to mine spatial-special information, auxiliary data sources like dark channel prior (Zhang et al., 2021) and geographic information (Chen et al., 2021) are also utilized in others. However, most deep learning-based cloud detection methods are conducted on mono-temporal images, lacking exploration of the temporal information, leading to poor performance in satellite videos. There is recently one method named TRCDNet (Luo et al., 2023) modeling spatial–temporal correspondence in multi-temporal geostationary satellite images, but the unexploited approach to gathering spatial–temporal information makes it insufficient in dense temporal data like sub-second-level satellite video.

2.2. CV video segmentation

Video segmentation aims to identify the key objects with specific properties or semantics in videos. In Zhou et al. (2022), video segmentation methods are reviewed thoroughly and classified into eight subclasses according to the input and output segmentation space. As the detected clouds have no instance-level information and the detection progress has no extra human inspection involvement, we regard automatic video object segmentation methods as the most related works to video cloud detection, where the input space refers to the video domain only without any manual initialization, along with the output space refers to a binary, foreground/background segmentation space without predefined semantic categories.

As no clear manual guidance is provided, the mainstream AVOS methods utilize motion information obtained from RGB images to identify the salient objects in a video, giving the inherent cognition that the objects and backgrounds have distinguished motions. Inspired by that, optical flows are calculated between adjacent frames. Together, motion and appearance features are fed into the network and fused to provide mutual guidance for object segmentation. In MATNet (Zhou et al., 2020), an asymmetric attention block was proposed for efficiently transforming appearance features into motion-related clues. In Isomer (Yuan et al., 2023), two transformer variants were designed to capture global contextual information and semantic correlation within motion and appearance features. To alleviate the influence of unstable motions, TMO (Cho et al., 2023) decoupled the feature fusion stage in encoders. Furthermore, DPA (Cho et al., 2024) introduced extra frames and an inter-frame attention module to provide a global context for the query frame. Although these algorithms can achieve promising performance in CV fields, the strong dependence on motion quality would decrease their performance when applied to satellite videos, as demonstrated in Section 1. A few other AVOS methods pay more attention to fully exploiting temporal coherence within a video, capturing recurring objects in the video for segmentation. Among these methods, COSNet (Lu et al., 2019) introduced a global co-attention mechanism to explore temporal dependence within a pair of frames. Wang et al. (2019) proposed an attentive graph neural network AGNN to model the relations between arbitrary frames. MED-VT (Karim et al., 2023) designed a multi-scale encoder–decoder transformer to extract spatial–temporal information implicitly. As clouds in satellite videos exhibit a fragmented and random distribution, global correspondence and spatial–temporal position modeling are of even greater importance in VCD tasks.

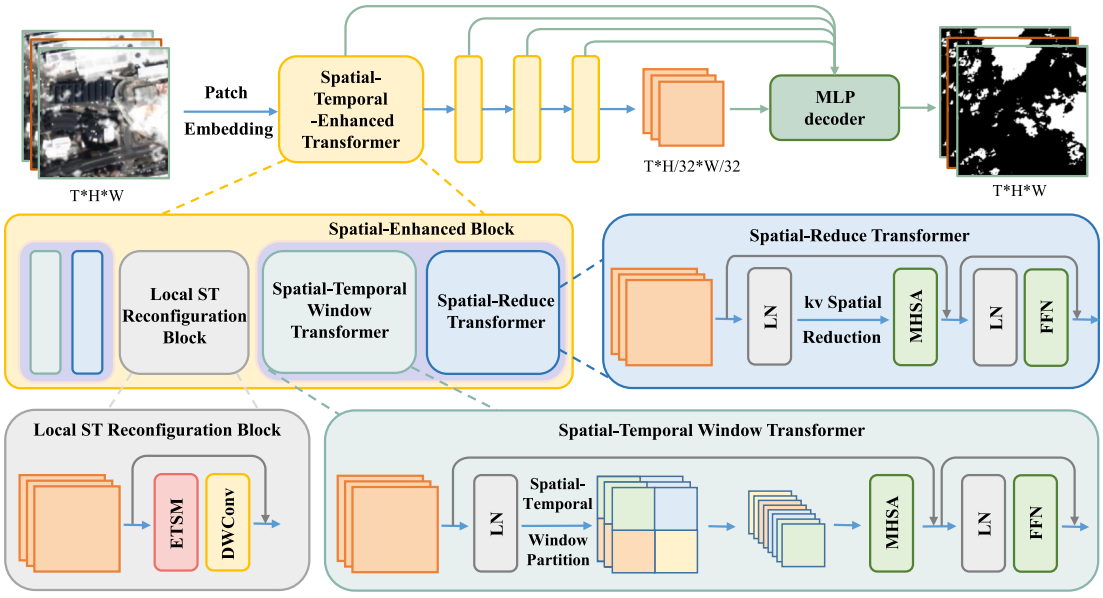


Fig. 3. Overall structure of our proposed network.

3. Methodology

3.1. Overview

We propose a transformer-based model VCDFormer to conduct effective cloud detection for remote sensing satellite videos. Given a video sequence with cloud cover $X_t \in R^{3\times H\times W} \mid t = 1, \ldots, T$, VCDFormer is designed to segment cloud pixels across frames, outputting a sequence of cloud masks $\hat{Y}_t \in R^{1\times H\times W}$. Taking the T-frame video sequence as input, VCDFormer performs feature encoding in a hierarchical architecture to capture multi-level information. Then the spatial–temporal-enhanced transformer is proposed to fully explore the inter-frame spatial–temporal information across multiple scales, which performs well in tackling the fragmented cloud and complex imaging interference. In STET, the leverage of the dynamic thin clouds property is strengthened by an effective module named local spatial–temporal reconfiguration block. Besides, the global correspondence is captured in a spatial-enhanced block, where both local and global spatial–temporal information can be utilized efficiently in a decoupled manner. Finally, VCDFormer employs a lightweight multilayer perceptron (MLP)-based decoder to aggregate the enhanced features from different levels and produce the final cloud mask sequence (Xie et al., 2021). Details are shown in Fig. 3.

3.2. Spatial-enhanced block

Clouds in satellite videos exhibit a fragmented and random distribution. Besides, the complicated imaging environment makes high reflections a challenging interference, which is indistinguishable from clouds and could lead to severe false detections. In that case, capturing global information might benefit the distinction between clouds and noises in the scenes, thereby improving the accuracy of the detection process. Transformer (Dosovitskiy, 2020) has been widely applied in modeling long-term dependency. Despite its remarkable capability of capturing global information, the extensive computational cost introduced by the inefficient self-attention mechanism is burdensome, preventing vanilla transformers from high-resolution tasks, including those in satellite videos. To conduct global attention modeling efficiently, we explore local and global spatial–temporal dependency in a decoupled manner. Additionally, to further enhance the global spatial–temporal correspondence extraction capability of the model, spatial-enhanced block is proposed. Specifically, SEB consists of a spatial–temporal window transformer and a spatial-reduce transformer, where the former models local-spatial and global-temporal information while the latter extracts global-spatial and local-temporal correspondence. By alternately applying these transformer blocks, both local and global information can be effectively modeled, thereby enhancing the ability to distinguish cloudy pixels from underlying surface interference.

3.2.1. Spatial–temporal window transformer

In transformer-based methods, window partition operation has been widely applied for images (Liu et al., 2021b; Li et al., 2022a), which aims at releasing the computation burden in high-resolution images. We extend the spatial window partition to a spatial–temporal window partition based on two key considerations. First, the high spatial–temporal resolution of satellite videos complicates the application of full attention across frames, leading to potential computational overload. Intuitively, partitioning the data into windows serves as an effective means of mitigating this issue. Additionally, transforming the spatial-window partition into a spatial–temporal one introduces no additional parameters, relying solely on a dimensional transformation while preserving the ability to model spatial–temporal correspondence within separate windows—an approach that meets the requirement for negligible computational cost. Second, the motion between adjacent frames in satellite videos is often minimal, typically on the order of a fraction of a pixel. The spatial–temporal window partition, which restricts self-attention calculation within the windows, enables the relative enlargement of motion within a specific spatial range, thereby emphasizing time-varying information.

Given an embedded feature $F_{\text{embed}} \in \mathbb{R}^{B\times T\times h\times w\times m}$, where h and w represent the number of embedded patches along the height and width dimensions, and m denoteds the feature dimension, LayerNorm (LN) is initially applied for feature normalization. The normalized feature is then passed into the spatial–temporal attention block, where spatial–temporal window partition and self-attention are performed. As illustrated in Fig. 3, the embedded feature is separated into local-spatial windows along the height and width dimensions using a window size ws. The resulting feature is reshaped into small spatial–temporal windows F_{stw}, upon which multi-head self-attention is applied to extract both local-spatial and global-temporal information. In each attention layer, linear layers are used to compute the query, key, and value

vectors from the spatial–temporal window features, followed by multi-head self-attention. Additionally, another LN and an MLP-based feed-forward network are utilized for feature projection. The process can be expressed as follows:

$$F_w \in R^{B\times T\times hg\times ws\times wg\times ws\times m} = \mathrm{STWP}(LN(F_{\mathrm{embed}}), ws) \tag{1}$$

$$F_{\mathrm{stw}} \in R^{B\times(hg\times wg)\times(T\times ws\times ws)\times m} = Reshape(F_w) \tag{2}$$

where STWP denotes the spatial–temporal window partition operation, and F_{stw} represents the reshaped small spatial–temporal windows. Linear layers are employed to compute the query, key, and value vectors (q, k, v), followed by the application of multi-head self-attention:

$$q, k, v = Linear(F_{stw}) \tag{3}$$

$$Attention(q, k, v) = Softmax(\frac{q(k)^T}{\sqrt{m}})v \tag{4}$$

$$F_{atten} = Linear(Reshape(Attention(q, k, v))) \tag{5}$$

3.2.2. Spatial-reduce transformer

In the spatial–temporal window transformer, while modeling the correspondence between local-spatial and global-temporal information, global information is often overlooked, leading to an inadequate capture of long-term dependencies. To address this issue, we incorporate a spatial-reduce transformer to model both local-temporal and global-spatial information more effectively (Xie et al., 2021). By alternately applying the spatial–temporal window transformer and the spatial-reduce transformer, we can efficiently explore both local and global spatial–temporal correspondences. The inclusion of global spatial context enhances the ability to distinguish slow-moving clouds and high-reflection areas, compensating for the limitations in local-spatial and global-temporal dependencies, and further improving the accuracy of cloud detection in complex imaging scenarios.

In the spatial-reduce transformer, we employ a stride convolution layer to aggregate information from multiple neighboring patches, forming a representation of the group features. This approach helps to alleviate the computational burden imposed by high spatial resolution. As depicted in Fig. 4, multi-head self-attention is applied to model the relationships between individual embedded patches (represented as single green windows) and the aggregated neighboring context patterns (depicted as grouped green windows). The combination of the spatial-reduce transformer and the hierarchical structure of VCDFormer allows the capture of multi-scale spatial dependencies (represented by yellow and green windows), which enables the integration of global spatial information and enhances the ability to distinguish between clouds and underlying surface noises. The operation of the spatial-reduce transformer can be formulated as follows:

$$F_s \in R^{BT\times h\times w\times m} = Reshape(LN(F_{embed})) \tag{6}$$

$$q = Linear(F_s) \tag{7}$$

$$k, v = Linear(StrideConv(F_s, scale)) \tag{8}$$

where StrideConv denotes the stride convolution layer to aggregate neighboring tokens with the downscale factor $scale$. Then multi-head self-attention following Eqs. (3)–(5) is performed using the original q and regrouped neighbor tokens k and v.

3.3. Local spatial–temporal reconfiguration block

As thin clouds in satellite videos often resemble land surface features, accurately detecting cloudy pixels in remote sensing scenes becomes challenging. Simply gathering spatial features like texture and context is inadequate. Traditional remote sensing methods resorted to the abundant information hidden in multiple bands, leveraging the spectral differences between clouds and surfaces to achieve good segmentation performance. As there is almost no redundant spectral information in RGB satellite videos, the utilization of temporal information becomes vital. Hence, we proposed an effective module named local spatial–temporal reconfiguration block in STEB to enhance the capability of capturing time-varying features and detecting dynamic thin clouds.

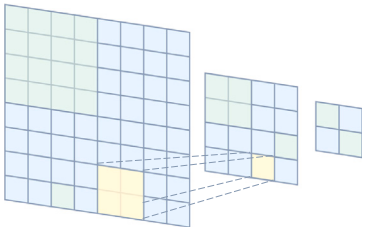


Fig. 4. The schematic of the spatial reduction operation in a hierarchical structure. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As shown in Fig. 3, the LSTRB consists of an enhanced temporal shift module (ETSM) and a depth-wise convolution. The proposed ETSM, as the core of our reconfiguration block, exchanges features in the temporal dimension. Denoting the input features as $F_{feat} \in R^{B\times T\times C\times H\times W}$ with C channels and T frames, ETSM shifts f channels forward and b channels backward at time dimension, as illustrated in Fig. 5. This parameter-free operation effectively reconfigures features across time, ensuring that each feature after ETSM contains a diverse multi-temporal combination across different channels. Following this, a 2D depth-wise convolution is applied to fuse local spatial–temporal features, with a residual connection added for improved learning. When T is set to 3, all ETSM-shifted features incorporate global temporal information, making the modeling of dynamic, time-varying features more efficient. Additionally, when the LSTRB is repeated across multiple stages, the temporal features are progressively regrouped in a cycle manner to capture a wider range of diverse feature reconfiguration patterns. This simple and lightweight LSTRB introduces negligible additional parameters while delivering significant performance improvements in video cloud detection tasks, particularly in exploring local spatial–temporal correspondence and accurately detecting dynamic thin clouds. The LSTRB progress can be formulated as:

$$[F_f, F_b, F_{remain}] = \mathrm{Split}(F_{feat}, b, f) \tag{9}$$

$$F_{f_s} = \text{Shift-f}(F_f) \tag{10}$$

$$F_{b_s} = \text{Shift-b}(F_b) \tag{11}$$

$$F_{ETSM} = [F_{f_s}, F_{b_s}, F_{remain}] \tag{12}$$

$$F_{LSTRB} = \mathrm{DWConv}(F_{ETSM}) + F_{ETSM} \tag{13}$$

where Split(·) denotes the channel split operation and Shift-f(·)/Shift-b(·) presents the forward/backward temporal shift operation. F_{LSTRB} denotes the reconfigure d features from LSTRB.

3.4. Loss function

We employ two loss functions to supervise the training of VCDFormer. The first one is the binary cross-entropy (BCE) loss measuring the difference between the centered frame of the predicted cloud mask sequences $\hat{Y}$ and the corresponding ground truth cloud mask Y, which is widely applied in image segmentation:

$$L_{bce}(Y, \hat{Y}) = -\frac{1}{N}\sum_{i=1}^{N} Y_i log(\hat{Y}_i) - (1 - Y_i) log(1 - \hat{Y}_i) \tag{14}$$

Since BCE loss provides pixel-wise constraints during training, to enforce region-wise supervision, we also incorporate intersection-over-union (IoU) loss. IoU loss quantifies the ratio of the intersection to

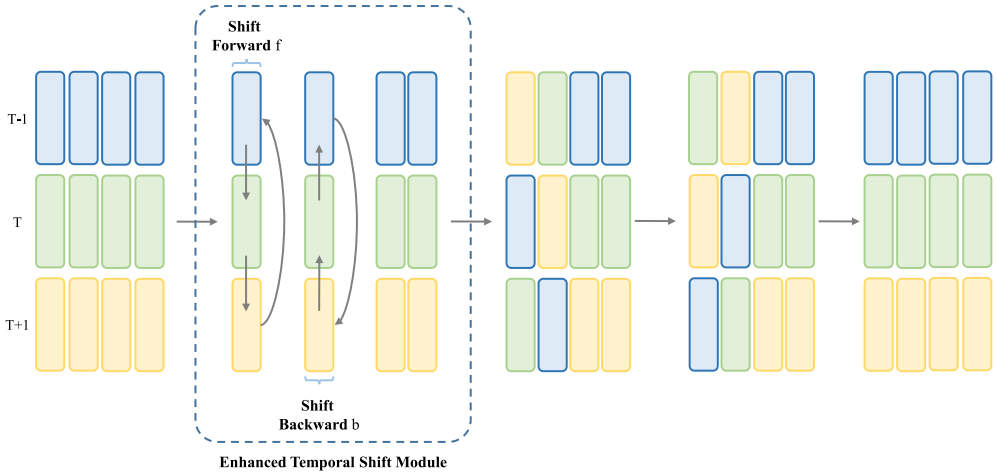


Fig. 5. The schematic of the enhanced temporal shift module and its cyclic process in multiple stages.

the union between the predicted and ground truth regions (Rezatofighi et al., 2019). The IoU loss is defined as:

$$L_{iou}(Y,\hat{Y}) = 1 - \frac{\sum_{i=1}^{N} Y_i \hat{Y}_i}{\sum_{i=1}^{N} Y_i + \hat{Y}_i - Y_i \hat{Y}_i} \tag{15}$$

The final loss function is formulated as a weighted combination of the two losses, with trade-off parameters λ_1 and λ_2 to balance their contributions:

$$Loss(Y,\hat{Y}) = \lambda_1 L_{bce}(Y,\hat{Y}) + \lambda_2 L_{iou}(Y,\hat{Y}) \tag{16}$$

4. Experiment

In this section, we first introduce the dataset generation process. Then our proposed VCDFormer is compared with several state-of-the-art methods (SOTA) methods from remote sensing and computer vision, along with ablation studies to validate the effectiveness of the proposed modules. Furthermore, comprehensive analyses are conducted to offer deeper insights into the proposed modules.

4.1. Dataset and experimental settings

4.1.1. WHU-VCD dataset

Existing remote sensing cloud detection datasets are either based on mono-temporal satellite images, or multi-temporal image sequences with an inherent revisit period, which is insufficient to provide dense dynamic information. Luo et al. (2023) proposed a multi-temporal dataset based on Fengyun4 images, whose temporal resolution is 15 min, far inferior to satellite videos. More importantly, the dataset is currently inaccessible, leaving research on dynamic satellite imagery largely underexplored. Another potential solution involves applying random mask generation algorithms in video inpainting methods (Li et al., 2022b). However, the randomly generated binary mask sequences lack cloud-related characteristics and can be easily segmented using a threshold, as illustrated in Fig. 7. To the best of our knowledge, no publicly available datasets have been specifically designed for VCD tasks. To effectively capture the motion characteristics of thick and thin clouds in sub-second-level satellite videos, we developed a synthetic dataset, WHU-VCD, using authentic cirrus cloud bands and satellite video scenes. As the first dataset specifically designed for video cloud detection, WHU-VCD will soon be made publicly available.[1]

The process of generating cloudy video sequences is exhibited in Fig. 6. We started by cropping cloud images from the cirrus band of Landsat 8. While this band is valuable for atmospheric correction and cloud masking, the clouds in these cropped images are static, making them unsuitable for simulating dynamic cloud motion in satellite videos. To simulate cloud movement, we used six parameters to move the clouds across frames. Two location parameters, h and w, determine the cloud positions within satellite video frames. Two motion parameters, m_h and m_w, control the speed and direction of cloud movement. These parameters are initialized randomly and adjusted throughout the process. We also introduced two threshold parameters for segmentation and scaling, which help limit cloud coverage within the video sequences and ensure visually pleasing results. Manual checks were performed at this stage to ensure proper visual effects and avoid overly thick or thin cloud scenes. Finally, the eligible cloud sequences were binarized to generate ground truth masks and combined with cloud-free frames from previous satellite video datasets (Xiao et al., 2021). Mean filtering is also applied at the edges of the cloudy areas to better mimic the characteristics of real-world clouds in satellite videos. Examples of synthetic cloudy scenes generated using various strategies, as well as real cloudy frames, are shown in Fig. 7. Overall, our strategy produces the most realistic cloud patterns.

We then applied this approach to generate the WHU-VCD dataset, which includes a training set of 189 cloudy scenes, each containing 100 frames. Additionally, an evaluation dataset was created with 18 videos, each containing 100 frames. This evaluation set utilizes six cloud-free videos from Jin et al. (2024) and 18 cloud masks to generate scenes with varying cloud coverage. The coverage is categorized into three classes: small coverage (S, ranging from 0.1–0.3), medium coverage (M, ranging from 0.35–0.6), and large coverage (L, ranging from 0.65–0.8), with each class containing 6 videos and a total of 600 images. With the WHU-VCD dataset, the motion characteristics of both thick and thin clouds in sub-second-level satellite videos can be effectively simulated.

4.1.2. Implement details

Our VCDFormer takes three frames randomly sampled and cropped to 512 × 512 as input. During the training process, random rotation and flipping for data augmentation are applied, and the batch size is set to 2 with a total of 150,000 training iterations. Adam optimizer with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are used to train our VCDFormer. The learning rate is initially set to 0.0002 and gradually decays according to a cosine annealing strategy. Experiments are conducted on the PyTorch framework with a single NVIDIA RTX 4090 GPU.

[1] The dataset is available at https://github.com/XyJin99/VCDFormer.

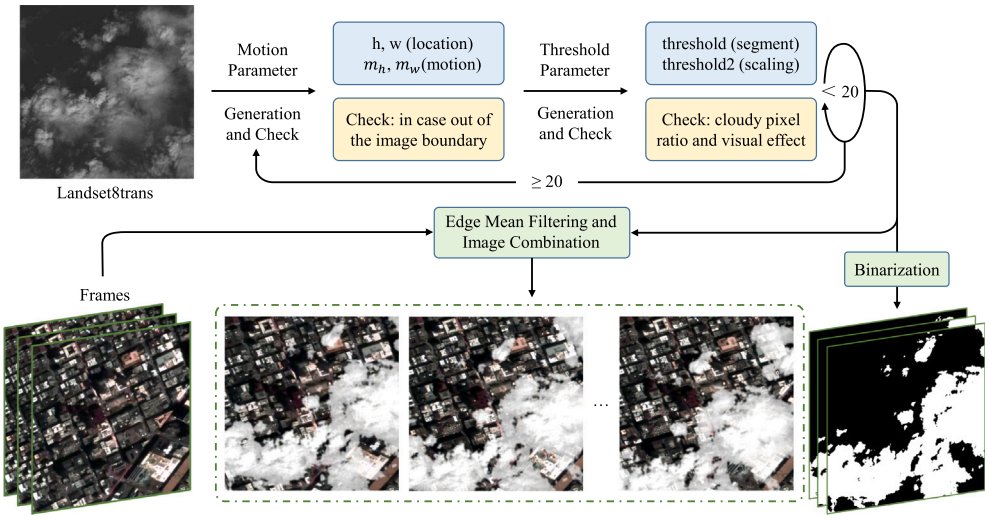


Fig. 6. The schematic of the simulated video cloud detection dataset process.

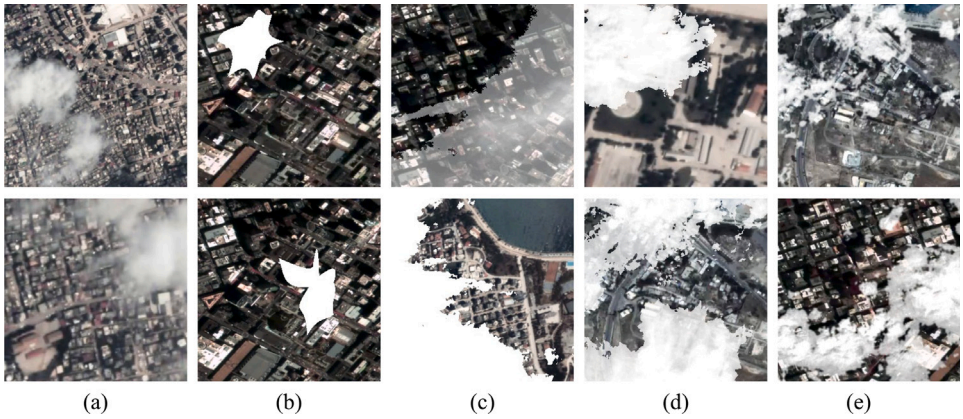


Fig. 7. Samples of synthetic cloudy scenes and real cloudy scenes. (a) Real cloudy satellite video frames. (b) Synthetic cloudy frames using CV video inpainting methods. (c) Synthetic cloudy frames using randomly generated threshold parameters without manual supervision. (d) Synthetic cloudy frames without edge mean filtering. (e) Synthetic cloudy frames in WHU-VCD.

4.2. Comparison with SOTA methods

We compared our proposed VCDFormer with eight state-of-the-art deep learning-based methods, including three remote sensing mono-temporal cloud detection methods, one modified multi-temporal cloud detection method, and four AVOS methods in the computer vision field. CDNetv2 (Guo et al., 2020), DCNet (Liu et al., 2021a), and BoundaryNets (Wu et al., 2022) are selected as RS cloud detection methods. TMO (Cho et al., 2023), MED-VT (Karim et al., 2023), DPA (Cho et al., 2024), and Isomer (Yuan et al., 2023) are selected as AVOS methods. Regarding the recently proposed multi-temporal cloud detection method TRCDNet (Luo et al., 2023), which does not have public code, we attempted to reproduce it following their paper. We also made adjustments to adapt it to the dense-temporal satellite video cloud detection task, which we denote as TRCDNet+. For fair comparisons, we carefully retrained those methods on our WHU-VCD dataset according to their original implementation. We choose both RS and CV metrics for a comprehensive comparison. Specifically, commonly used remote sensing metrics, including Precision, Recall, IoU, Overall Accuracy (OA), and F1-Score are selected for evaluation. Standard evaluation metrics for AVOS task, including the mean of region similarity J, mean of contour accuracy F, and $J\&F$ average, are presented as CV metrics (Perazzi et al., 2016). It is worth noting that J is equal to a normal intersection-over-union metric. We include both metrics here to ensure the completeness of the evaluation systems in both RS and CV fields.

4.2.1. Quantitative results

Quantitative comparison of our WHU-VCD evaluation dataset is reported in Table 1. Mono-temporal RS cloud detection methods like DCNet and CDNetv2 produce poor detection results, which is mainly caused by the lack of utilization of temporal information. The limited receptive field introduced by their convolution-based modules also prevents them from better performance. Paying more attention to the multi-scale features and modeling the boundary details explicitly, BoundaryNets outperforms other mono-temporal-based methods, showing a remarkable improvement in the precision and IoU metrics. However, BoundaryNets still shows inferior performance on recall, which indicates that miss detections arise frequently in results. As for the multi-temporal cloud detection method, TRCDNet+ shows improved performance over DCNet and CDNetv2, yet shows even worse detection accuracy than BoundaryNets. We speculate that the lack of near-infrared spectra and the simplistic approach to capturing spatio-temporal information may hinder its ability to handle dense-temporal data like satellite videos. For AVOS methods, motion-based approaches such as TMO and DPA exhibit inferior detection performance compared to MED-VT, which may be attributed to the unstable quality of adjacent optical flows. On the contrary, MED-VT performs video object segmentation within multiple frames without motion cues, which results in superior effectiveness. Besides, Isomer fuses appearance and motion features with channel attention, implicitly mitigating the impact of unreliable motion information and producing more stable results. Overall,

Table 1
Quantitative comparisons on our WHU-VCD dataset. The **best** and second-best results are highlighted.

Models	Coverage	RS metrics					CV metrics		
		Precision	Recall	IoU	OA	F1-Score	J	F	$J\&F$
DCNet	S	0.8168	0.8493	0.7180	0.9335	0.8276	0.7180	0.7217	0.7198
	M	0.9024	0.9213	0.8393	0.9331	0.9110	0.8393	0.7818	0.8105
	L	0.9641	0.9792	0.9448	0.9577	0.9716	0.9448	0.8481	0.8965
	Aver	0.8944	0.9166	0.8340	0.9414	0.9034	0.8340	0.7839	0.8090
CDNetv2	S	0.8931	0.7642	0.7048	0.9382	0.8177	0.7048	0.7950	0.7499
	M	0.9548	0.8644	0.8301	0.9328	0.9058	0.8301	0.8034	0.8167
	L	0.9856	0.9464	0.9336	0.9501	0.9655	0.9336	0.8786	0.9061
	Aver	0.9445	0.8583	0.8228	0.9404	0.8963	0.8228	0.8257	0.8242
TRCDNet+	S	0.8727	0.8494	0.7613	0.9487	0.8589	0.7613	0.8304	0.7959
	M	0.9498	0.8959	0.8568	0.9433	0.9218	0.8568	0.8572	0.8570
	L	0.9821	0.9616	0.9451	0.9586	0.9717	0.9451	0.9051	0.9251
	Aver	0.9349	0.9023	0.8544	0.9502	0.9175	0.8544	0.8642	0.8593
BoundaryNets	S	0.9233	0.8453	0.7934	0.9586	0.8808	0.7934	0.8869	0.8401
	M	0.9616	0.9040	0.8733	0.9502	0.9316	0.8733	0.8687	0.8710
	L	0.9831	0.9674	0.9516	0.9635	0.9752	0.9516	0.9104	0.9310
	Aver	0.9560	0.9056	0.8728	0.9574	0.9292	0.8728	0.8887	0.8807
TMO	S	0.8783	0.8470	0.7629	0.9475	0.8590	0.7629	0.8431	0.8030
	M	0.9429	0.9092	0.8623	0.9450	0.9252	0.8623	0.8538	0.8581
	L	0.9837	0.9634	0.9483	0.9610	0.9734	0.9483	0.9085	0.9284
	Aver	0.9350	0.9065	0.8578	0.9512	0.9192	0.8578	0.8685	0.8631
MED-VT	S	0.8965	0.9043	0.8229	0.9645	0.9001	0.8229	0.9196	0.8713
	M	0.9434	0.9256	0.8780	0.9506	0.9343	0.8780	0.8790	0.8785
	L	0.9827	0.9645	0.9484	0.9608	0.9734	0.9484	0.8957	0.9220
	Aver	0.9409	0.9315	0.8831	0.9586	0.9360	0.8831	0.8981	0.8906
DPA	S	0.9086	0.8560	0.7930	0.9589	0.8801	0.7930	0.8832	0.8381
	M	0.9325	0.9288	0.8717	0.9481	0.9305	0.8717	0.8819	0.8768
	L	0.9749	0.9715	0.9479	0.9603	0.9732	0.9479	0.9173	0.9326
	Aver	0.9387	0.9188	0.8708	0.9558	0.9279	0.8708	0.8941	0.8825
Isomer	S	0.9419	0.9305	0.8823	0.9774	0.9360	0.8823	0.9499	0.9161
	M	0.9632	0.9550	0.9220	0.9695	0.9590	0.9220	0.9399	0.9309
	L	0.9863	0.9843	0.9710	0.9781	0.9853	0.9710	0.9557	0.9634
	Aver	0.9638	0.9566	0.9251	0.9750	0.9601	0.9251	0.9485	0.9368
Ours	S	**0.9561**	**0.9675**	**0.9274**	**0.9864**	**0.9617**	**0.9274**	**0.9812**	**0.9543**
	M	**0.9729**	**0.9752**	**0.9498**	**0.9805**	**0.9741**	**0.9498**	**0.9738**	**0.9618**
	L	**0.9901**	**0.9921**	**0.9824**	**0.9867**	**0.9911**	**0.9824**	**0.9796**	**0.9810**
	Aver	**0.9731**	**0.9783**	**0.9532**	**0.9845**	**0.9756**	**0.9532**	**0.9782**	**0.9657**

VCDFormer, with its design tailored to the characteristics of satellite videos, demonstrates outstanding performance in video cloud detection tasks. It outperforms all other methods across eight evaluation metrics, regardless of the cloud coverage in the scenes.

4.2.2. Qualitative results

Visual results from six video scenes are selected for qualitative comparison. To enhance visualization, miss-detection areas are marked in green, while false-detection areas are highlighted in red. As illustrated in Fig. 8, mono-temporal cloud detection methods, such as DCNet and CDNetv2, perform poorly in detecting cloudy pixels within frames. TRCDNet+ and BoundaryNets demonstrate slight improvements but still fail to achieve accurate detection in complex scenes. On average, AVOS algorithms outperform remote sensing methods, which is understandable. Remote sensing methods primarily rely on spectral-spatial information for cloud detection, while AVOS methods leverage spatio-temporal features for object segmentation. When applied to sub-second-level satellite videos with only three basic channels, RS methods lose access to spectral information like near-infrared bands, which is critical for capturing cloud properties across the spectral dimension. Consequently, the performance of remote sensing methods suffers in this context.

Among AVOS methods, TMO exhibits notable false detections, while DPA produces cloud masks dominated by miss-detection areas. MED-VT and Isomer achieve better average results but still leave room for improvement, particularly around the edges of cloudy areas. In comparison, our proposed VCDFormer demonstrates remarkable performance in detecting dynamic cloudy pixels across various levels of cloud coverage, producing the fewest miss and false detections in diverse scenes.

4.2.3. Real-world results

Real-world experiments were conducted on Jilin-1 satellite videos to evaluate the generalization ability of the proposed model. To ensure consistency with prior studies on video cloud removal methods, two real-world scenes are presented in Figs. 9 and 10, along with the corresponding cloud masks generated by various methods for comparison.

Our proposed VCDFormer demonstrates significant improvements in dynamic cloud detection within real-world scenarios. While most comparison methods struggle, particularly in identifying thin cloudy areas that are challenging to distinguish, VCDFormer consistently delivers superior results. It produces stable detection outcomes by accurately segmenting cloudy pixels from complex backgrounds, even in difficult conditions.

Notably, we employed a cross-frame sampling strategy in the real-world experiments, which samples reference frames at intervals to capture more time-varying information within the sequence. This intuitive and cost-free approach enhances the exploration of spatio-temporal information through the input data and contributes to the stable detection performance of VCDFormer.

4.3. Ablation study

4.3.1. Spatial-enhanced block

Vanilla vision transformer models capture long-term dependencies in the spatial dimension of images, but the huge computational cost

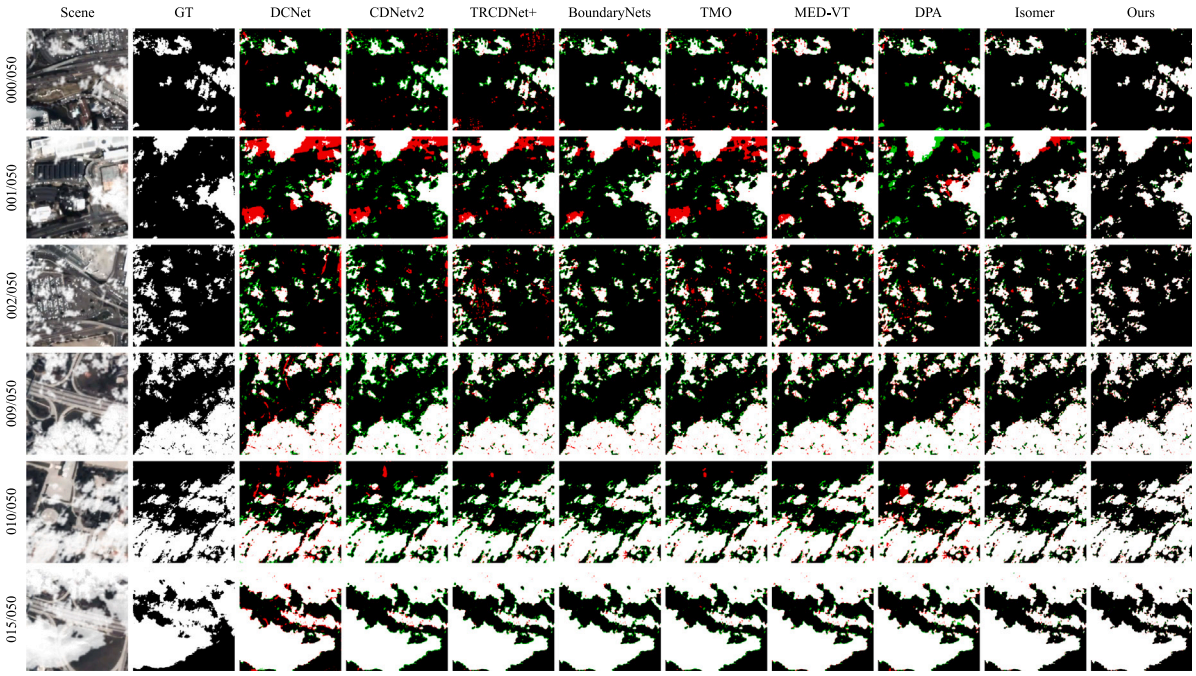


Fig. 8. Visual comparisons with SOTA models on the WHU-VCD dataset including three small cloud coverage scenes (000/001/002), two medium coverage scenes (009/010), and one large coverage scene (015). We denote the miss detections areas in green and the false detections in red for better visualization. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Ablation studies of SEB and LSTRB conducted on three 512 × 512 frames. *win_st* and *sr* denote the utilized proposed spatial–temporal window transformer and spatial-reduce transformer, respectively. $Swin$ and Sep indicate the extended 3D version of Swin-transformer (Liu et al., 2021b). $3DDWC$ and *ST_token* represent the 3D depth-wise convolution and the spatial–temporal position embedding in Li et al. (2025). The best results in the ablation study of SEB and LSTRB are highlighted in underlined and **bold**, respectively.

LSTRB	Attention	Precision	Recall	IoU	OA	F1	J	F	$J\&F$
–	Spatial	Out of memory							
–	Spatial & temporal	Out of memory							
–	*win_st*	0.9658	0.9673	0.9366	0.9791	0.9664	0.9366	0.9594	0.9480
–	*sr*	0.9497	0.9602	0.9159	0.9715	0.9543	0.9159	0.9331	0.9245
–	*win_st+sr* (SEB)	0.9653	0.9697	0.9383	0.9797	0.9673	0.9383	0.9652	0.9517
–	*win_st+Swin*	0.9668	0.9675	0.9378	0.9796	0.9671	0.9378	0.9608	0.9493
–	*win_st+Sep*	0.9613	0.9688	0.9340	0.9782	0.9649	0.9340	0.9580	0.9460
✓	*win_st*	0.9703	**0.9784**	0.9507	0.9836	0.9743	0.9507	0.9768	0.9637
✓	*sr*	0.9698	0.9760	0.9480	0.9827	0.9728	0.9480	0.9747	0.9614
✓	SEB	**0.9731**	0.9783	**0.9532**	**0.9845**	**0.9756**	**0.9532**	**0.9782**	**0.9657**
$3DDWC$	SEB	0.9696	0.9723	0.9446	0.9820	0.9709	0.9446	0.9688	0.9567
ST_token	SEB	0.9671	0.9722	0.9422	0.9809	0.9695	0.9422	0.9689	0.9555

makes it inefficient for high temporal–spatial resolution satellite videos. As shown in Table 2, when trained on a three-frame input at a resolution of 512 × 512, applying self-attention to both spatial–temporal and spatial dimensions failed due to insufficient GPU memory.

By implementing the spatial-reduce transformer and spatial–temporal window transformer, spatial–temporal information is modeled in a local–global manner, which shows effectiveness but still falls short in capturing the long-term spatial–temporal correspondence. To address this, we combine both transformer blocks, enabling the decoupled modeling of global spatial–temporal dependencies, which leads to a further improvement in model performance. This combination forms our spatial-enhanced block. Additionally, we extend two window-based transformer blocks (Liu et al., 2021b; Li et al., 2022a) designed for image tasks to a 3D spatial–temporal architecture for comparison. Overall, SEB delivers better detection results than the other methods, demonstrating the effectiveness of our proposed block.

4.3.2. Local spatial–temporal reconfiguration block

To further enhance the model's ability to capture dynamic clouds, we introduced the simple yet effective LSTRB module. As shown in Table 2, the implementation of LSTRB results in a significant performance improvement across various transformer blocks. Moreover, replacing LSTRB with 3D depth-wise convolution decreases performance, as it lacks the explicit temporal reconfiguration information provided by the ETSM. We also utilized a lightweight spatial–temporal position embedding introduced in Li et al. (2025) for comparison. However, the fixed size of spatial–temporal tokens is inflexible in handling varying resolutions, resulting in even worse performance than 3D depth-wise convolution. In contrast, our LSTRB is flexible across different resolutions and consistently produces accurate detection results.

4.4. Discussion

In this section, we further elaborate on several conclusions previously mentioned to provide a deeper understanding of the modules employed in our approach. We also discuss the limitations of our method and potential future directions for improvement.

4.4.1. Robust detection in complex imaging environments

As noted in Section 1, thin clouds and high-reflection areas significantly complicate the imaging process, posing additional challenges

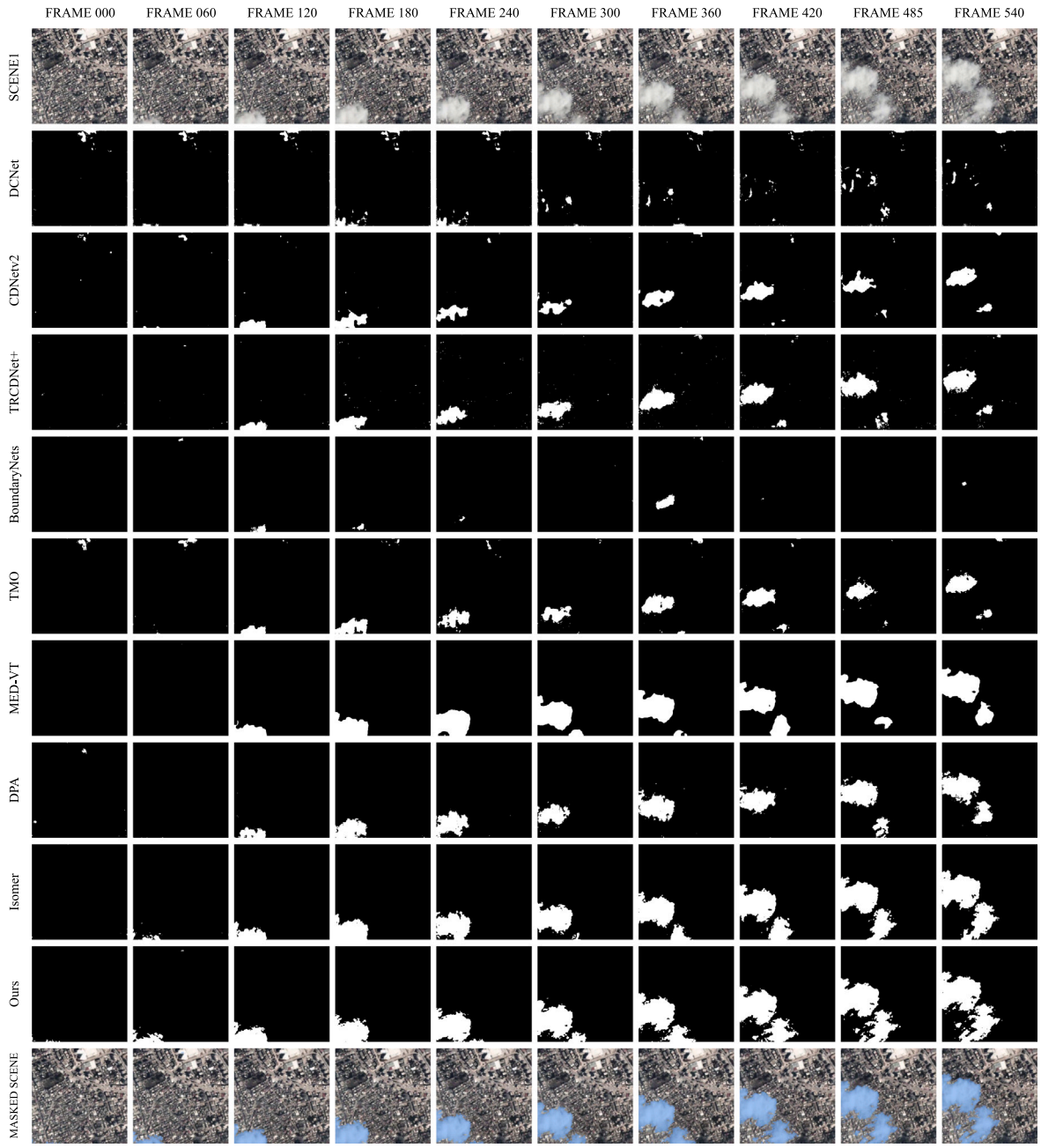


Fig. 9. Real-world comparisons on SCENE1 with SOTA models.

in modeling spatial–temporal information in satellite videos. This observation leads to a critical question: *can the algorithm effectively leverage spatial–temporal information to distinguish clouds from complex backgrounds, such as high-reflection areas?*

To evaluate the robustness of various methods in extremely complicated backgrounds, we selected six scenes from the WHU-VCD evaluation dataset, whose backgrounds include three cloud-free videos with dynamic or static high-reflection areas. We compared the previously discussed RS and CV methods in these scenes. As shown in Fig. 11, RS mono-temporal detection methods struggle with correctly segmenting clouds from static high-reflections, such as building roofs in scenes 001 and 007. Additionally, small and fragmented high-reflection areas in scenes 000 and 002 were falsely detected as clouds by CDNetv2. This issue could be mitigated by the deformable and multi-scale features modeled in DCNet and BoundaryNets. Besides, TRCDNet+ tends to misclassify inter-frame glint as clouds, due to the similarity in time-varying properties between glint and cloudy pixels. AVOS methods generally perform better, as expected, because they utilize motion information to suppress interference from blinking areas. Nevertheless, The unstable optical flow leads to additional miss or false detections at the boundary, as observed in the DPA and Isomer results. In contrast, VCDFormer is specifically designed to handle the unique characteristics of satellite videos, discarding explicit motion cue modeling and fully exploring multi-scale global spatial–temporal correspondence. Overall, VCDFormer demonstrates robust performance within extremely complex imaging environments.

4.4.2. *Effectiveness of LSTRB and SEB in real-world scenes*

Additional visual results are provided to offer an intuitive understanding of the effectiveness of our proposed modules. As described in Fig. 12, static building roofs are miss-detected and thin clouds go unrecognized without LSTRB, highlighting its crucial role in improving detection accuracy. Moreover, in Fig. 13, algorithms lacking

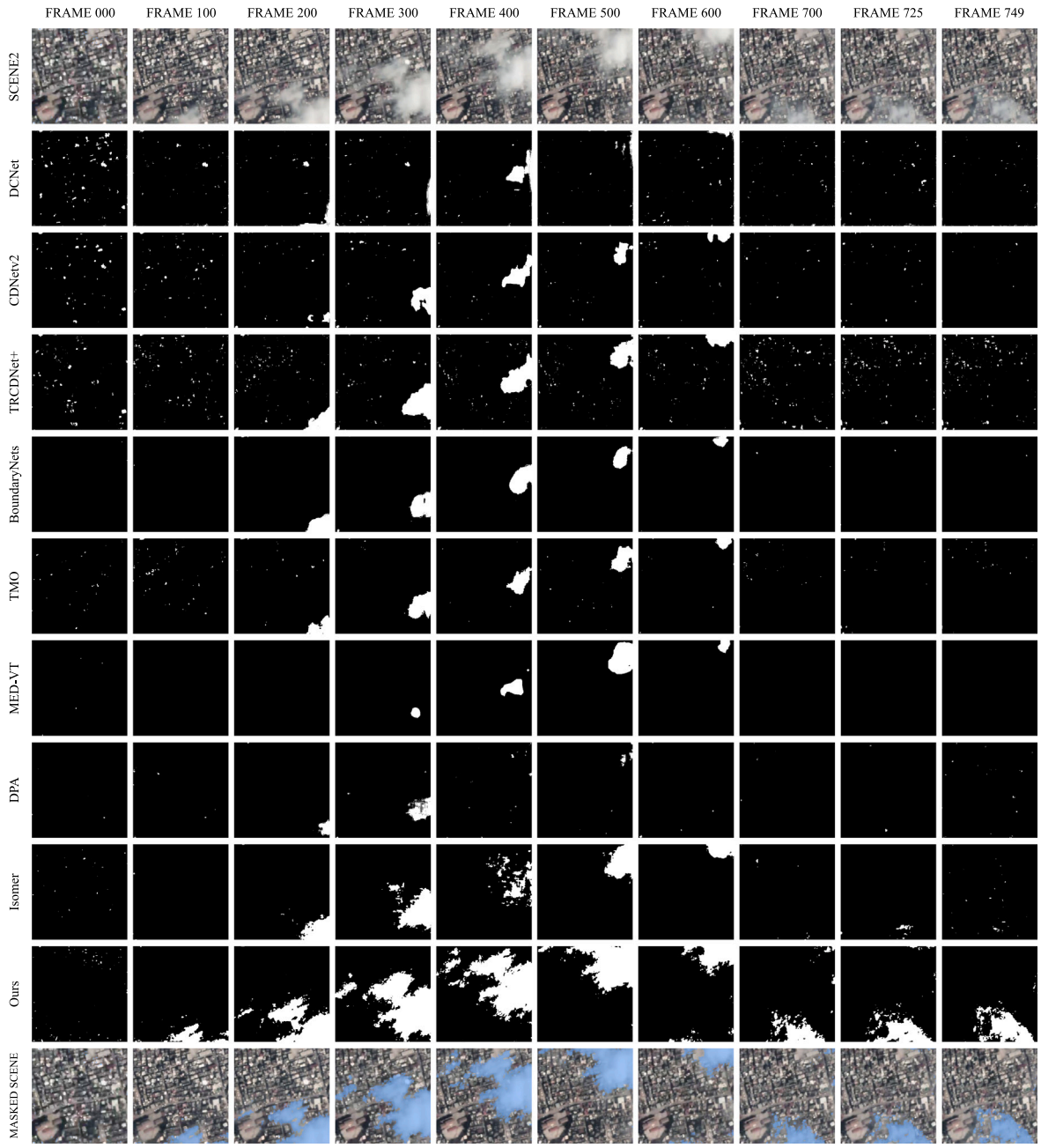


Fig. 10. Real-world comparisons on SCENE2 with SOTA models.

either the spatial–temporal window transformer or spatial-reduce transformer suffer from severe false detections in cloud-free frames and produce fragmented segmentation within cloudy areas. In contrast, by alternately applying these two transformers, VCDFormer effectively detects cloudy pixels and produces visually coherent results, accurately distinguishing clouds from the background.

4.4.3. Video cloud removal with predicted cloud masks

Accurate cloud detection is essential for effective video cloud removal, enabling seamless reconstruction of the Earth's surface, as demonstrated in our previous work (Jin et al., 2024). By combining VCDFormer with RFE-VCR, we establish a pipeline capable of automatically detecting and removing clouds in satellite videos, providing a clear, long-term view of the Earth's surface. For consistency with our prior study, we use the detected cloud masks to guide the satellite video cloud removal task. Cloud masks predicted by MED-VT, Isomer, and VCDFormer, are selected to guide the RFE-VCR process. As shown in Fig. 14, the cloud removal performance improves as the precision of the cloud masks increases. The comparative methods fail to properly detect large areas of thin clouds, which results in unsatisfactory removal performance. In contrast, the accurate cloud masks produced by VCDFormer effectively aid the cloud removal process, yielding more visually appealing results and providing indirect evidence of the effectiveness of our approach.

4.4.4. Limitations and future improvements

As the first method for sub-second-level satellite video cloud detection, there are still some limitations that can be addressed in future work. For instance, in this study, we treat high-reflection areas as interference when detecting thin and thick clouds. However, in practical applications, these areas should be identified and excluded, such as high reflections from water or building roofs. Currently, accurately

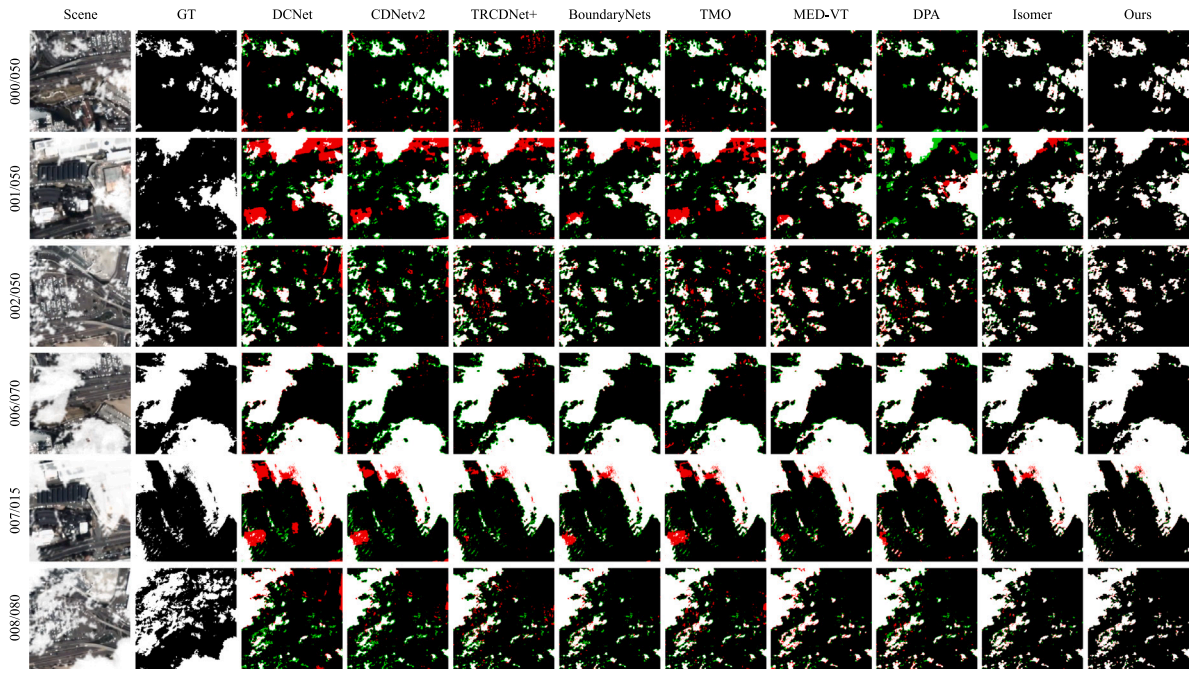


Fig. 11. Visual comparisons with SOTA models on six scenes of WHU-VCD evaluation dataset with dynamic or static high-reflection areas.

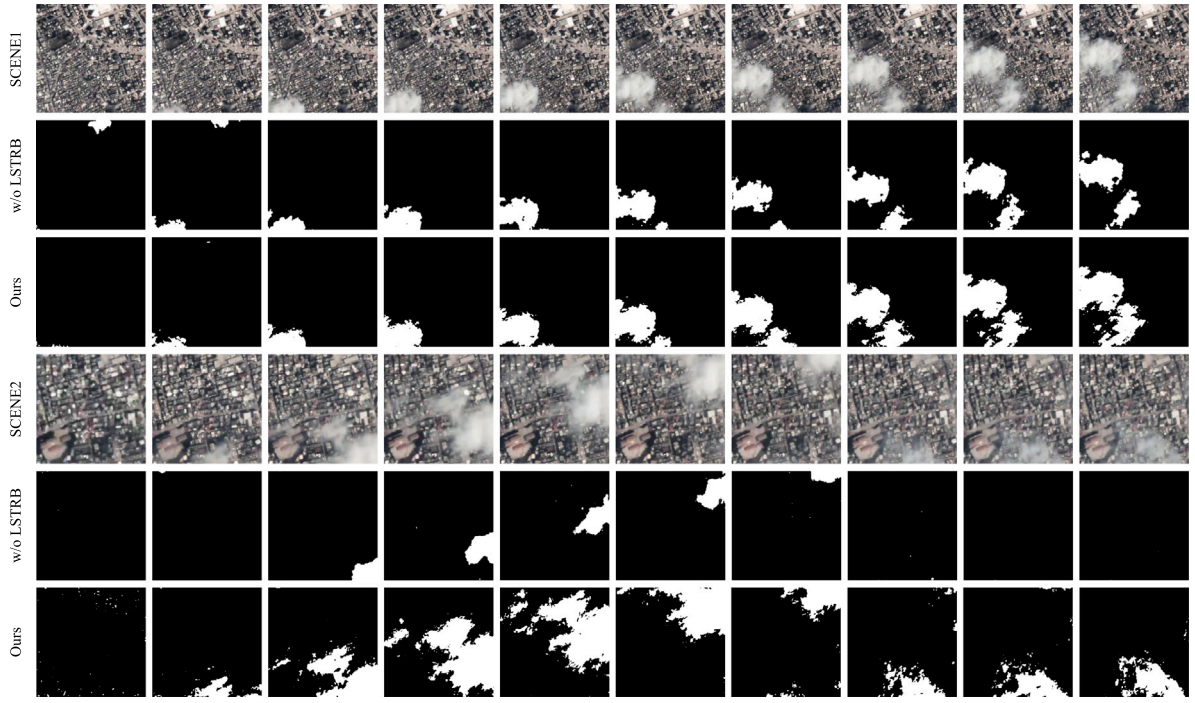


Fig. 12. Predicted masks with/without LSTRB in real-world scenarios.

distinguishing these reflections from cloud pixels remains challenging due to a lack of training data to simulate the dynamic scaling characteristics of high-reflection areas. A potential solution could involve domain generalization or domain adaptation strategies, which extend the video cloud detection task to handle time-varying information, similar to video semantic segmentation tasks.

Future improvements can be categorized into three key directions:

1. **Network development**: Advancing the architecture, such as integrating models like Mamba and its variants, could allow for more efficient long-term spatial–temporal exploration. This may offer a promising alternative to transformers for capturing long-range dependencies in satellite video data.
2. **Task extension**: Expanding the scope of video cloud detection tasks to jointly address time-varying information detection would enhance the ability to capture dynamic environmental changes and further improve cloud detection accuracy.
3. **Multi-task collaboration**: Cloud detection and cloud removal are closely related tasks. Accurate cloud detection masks can guide cloud removal processes, and the resulting cloud removal outcomes can, in turn, refine cloud detection. Treating video cloud detection and cloud removal as interrelated tasks that

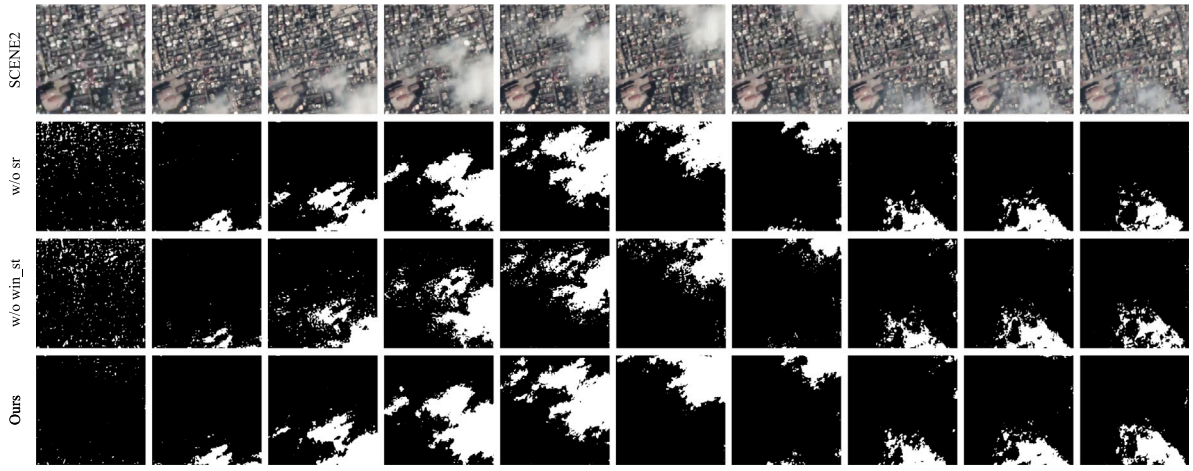


Fig. 13. Predicted masks with/without SEB in real-world scenarios.

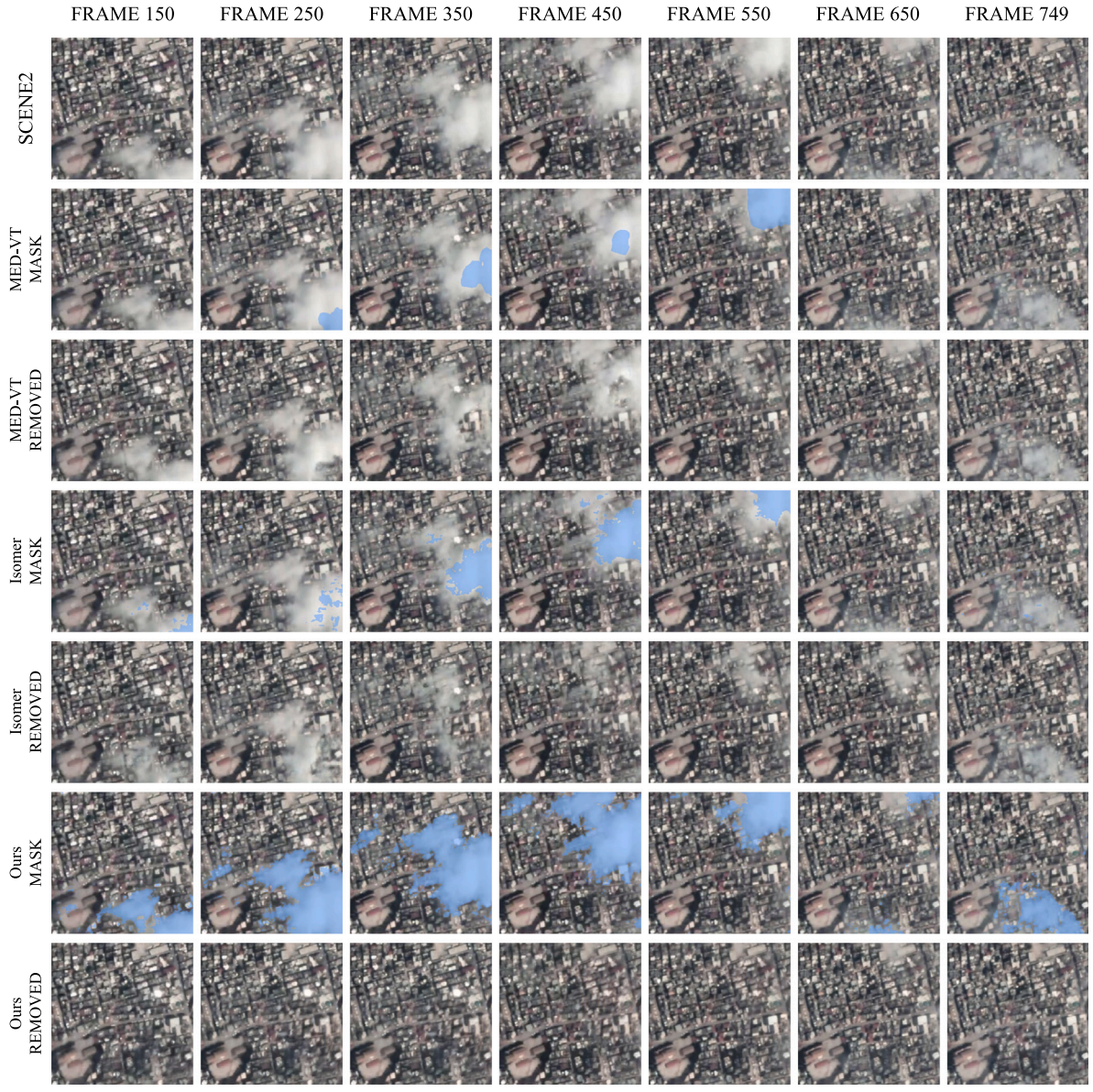


Fig. 14. Video cloud removal results compared with SOTA methods.

can be jointly optimized holds great potential. We are actively working on this approach.

5. Conclusion

Cloud occlusion is a common atmospheric phenomenon, and accurately capturing its motion and location is crucial for utilizing satellite videos to achieve continuous surface monitoring. In this paper, we propose VCDFormer, the first sub-second-level video cloud detection framework. VCDFormer employs a Spatial-Temporal-Enhanced Transformer that effectively captures inter-frame spatial–temporal dependencies across multiple scales. Two novel modules, the Local Spatial-Temporal Reconfiguration Block and Spatial-Enhanced Block, are utilized to adeptly explore both local and global spatial–temporal information, significantly improving dynamic cloud detection accuracy while reducing interference from complex imaging conditions. To facilitate this research, we have also developed the WHU-VCD dataset, the first sub-second-level synthetic dataset for satellite video cloud detection. Our experiments underscore the robustness and effectiveness of VCDFormer. Future work will focus on refining the model and exploring new solutions to enhance the quality of satellite video cloud detection.

CRediT authorship contribution statement

Xianyu Jin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jiang He:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Yi Xiao:** Writing – review & editing, Resources, Funding acquisition. **Ziyang Lihe:** Writing – review & editing, Investigation. **Jie Li:** Supervision. **Qiangqiang Yuan:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China 42230108, 42471414, 423B2104, and 424B2009.

Data availability

Data will be made available on request.

References

An, Z., Shi, Z., 2015. Scene learning for cloud detection on remote-sensing images. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 8 (8), 4206–4222.

Braaten, J.D., Cohen, W.B., Yang, Z., 2015. Automated cloud and cloud shadow identification in Landsat MSS imagery for temperate ecosystems. Remote Sens. Environ. 169, 128–138.

Chai, D., Newsam, S., Zhang, H.K., Qiu, Y., Huang, J., 2019. Cloud and cloud shadow detection in landsat imagery based on deep convolutional neural networks. Remote Sens. Environ. 225, 307–316.

Chen, Y., Tang, Y., Xiao, Y., Yuan, Q., Zhang, Y., Liu, F., He, J., Zhang, L., 2024. Satellite video single object tracking: A systematic review and an oriented object tracking benchmark. ISPRS- J. Photogramm. Remote. Sens. 210, 212–240.

Chen, Y., Weng, Q., Tang, L., Liu, Q., Fan, R., 2021. An automatic cloud detection neural network for high-resolution remote sensing imagery with cloud–snow coexistence. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.

Cho, S., Lee, M., Lee, S., Lee, D., Choi, H., Kim, I.-J., Lee, S., 2024. Dual prototype attention for unsupervised video object segmentation. In: Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit. pp. 19238–19247.

Cho, S., Lee, M., Lee, S., Park, C., Kim, D., Lee, S., 2023. Treating motion as option to reduce motion dependency in unsupervised video object segmentation. In: Proc IEEE Winter Conf. Appl. Comput. Vis.. pp. 5140–5149.

Deng, C., Li, Z., Wang, W., Wang, S., Tang, L., Bovik, A.C., 2018. Cloud detection in satellite images based on natural scene statistics and gabor features. IEEE Geosci. Remote. Sens. Lett. 16 (4), 608–612.

Dong, J., Wang, Y., Yang, Y., Yang, M., Chen, J., 2024. MCDNet: Multilevel cloud detection network for remote sensing images based on dual-perspective change-guided and multi-scale feature fusion. Int. J. Appl. Earth Obs. Geoinf. 129, 103820.

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR - Int. Conf. Learn. Represent..

Francis, A., Sidiropoulos, P., Muller, J.-P., 2019. CloudFCN: Accurate and robust cloud detection for satellite imagery with deep learning. Remote. Sens. 11 (19), 2312.

Ghasemian, N., Akhoondzadeh, M., 2018. Introducing two Random Forest based methods for cloud detection in remote sensing images. Adv. Space Res. 62 (2), 288–303.

Guo, J., Yang, J., Yue, H., Tan, H., Hou, C., Li, K., 2020. CDnetV2: CNN-based cloud detection for remote sensing imagery with cloud-snow coexistence. IEEE Trans. Geosci. Remote Sens. 59 (1), 700–713.

He, J., Yuan, Q., Li, J., Xiao, Y., Liu, D., Shen, H., Zhang, L., 2023a. Spectral super-resolution meets deep learning: achievements and challenges. Inf. Fusion 97, 101812.

He, J., Yuan, Q., Li, J., Xiao, Y., Zhang, L., 2023b. A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection. ISPRS- J. Photogramm. Remote. Sens. 204, 131–144.

Hollstein, A., Segl, K., Guanter, L., Brell, M., Enesco, M., 2016. Ready-to-use methods for the detection of clouds, cirrus, snow, shadow, water and clear sky pixels in sentinel-2 MSI images. Remote. Sens. 8 (8), 666.

Irish, R.R., Barker, J.L., Goward, S.N., Arvidson, T., 2006. Characterization of the landsat-7 ETM+ automated cloud-cover assessment (ACCA) algorithm. Photogramm. Eng. Remote. Sens. 72 (10), 1179–1188.

Jin, X., He, J., Xiao, Y., Lihe, Z., Liao, X., Li, J., Yuan, Q., 2024. RFE-VCR: Reference-enhanced transformer for remote sensing video cloud removal. ISPRS-J. Photogramm. Remote. Sens. 214, 179–192.

Karim, R., Zhao, H., Wildes, R.P., Siam, M., 2023. MED-VT: Multiscale encoder-decoder video transformer with application to object segmentation. In: Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit. pp. 6323–6333.

Li, K., Li, X., Wang, Y., He, Y., Wang, Y., Wang, L., Qiao, Y., 2025. Videomamba: State space model for efficient video understanding. In: ArXiv. Springer, pp. 237–255.

Li, Z., Lu, C.-Z., Qin, J., Guo, C.-L., Cheng, M.-M., 2022b. Towards an end-to-end framework for flow-guided video inpainting. In: Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit. pp. 17562–17571.

Li, J., Wang, Q., 2024. CSDFormer: A cloud and shadow detection method for landsat images based on transformer. Int. J. Appl. Earth Obs. Geoinf. 129, 103799.

Li, W., Wang, X., Xia, X., Wu, J., Li, J., Xiao, X., Zheng, M., Wen, S., 2022a. Sepvit: Separable vision transformer. ArXiv.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021b. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc IEEE Int Conf Comput Vision. pp. 10012–10022.

Liu, Y., Wang, W., Li, Q., Min, M., Yao, Z., 2021a. DCNet: A deformable convolutional cloud detection network for remote sensing imagery. IEEE Geosci. Remote. Sens. Lett. 19, 1–5.

Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F., 2019. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit. pp. 3623–3632.

Luo, C., Feng, S., Quan, Y., Ye, Y., Li, X., Xu, Y., Zhang, B., Chen, Z., 2023. TRCDNet: A transformer network for video cloud detection. IEEE Trans. Geosci. Remote Sens..

Mateo-García, G., Adsuara, J.E., Pérez-Suay, A., Gómez-Chova, L., 2019. Convolutional long short-term memory network for multitemporal cloud detection over landmarks. In: IGARSS 2019-2019. IEEE, pp. 210–213.

Mohajerani, S., Saeedi, P., 2019. Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery. In: Dig Int Geosci Remote Sens Symp. IGARSS, IEEE, pp. 1029–1032.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A., 2016. A benchmark dataset and evaluation methodology for video object segmentation. In: Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit. pp. 724–732.

Qiu, S., Zhu, Z., Woodcock, C.E., 2020. Cirrus clouds that adversely affect landsat 8 images: What are they and how to detect them? Remote Sens. Environ. 246, 111884.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In: Proc IEEE Comput Soc Conf Comput Vision Pattern Recognit. pp. 658–666.

Sun, L., Wei, J., Wang, J., Mi, X., Guo, Y., Lv, Y., Yang, Y., Gan, P., Zhou, X., Jia, C., et al., 2016. A universal dynamic threshold cloud detection algorithm (UDTCDA) supported by a prior surface reflectance database. J. Geophys. Res. Atmos. 121 (12), 7172–7196.

Wang, W., Lu, X., Shen, J., Crandall, D.J., Shao, L., 2019. Zero-shot video object segmentation via attentive graph neural networks. In: Proc IEEE Int Conf Comput Vision. pp. 9236–9245.

Wu, K., Xu, Z., Lyu, X., Ren, P., 2022. Cloud detection with boundary nets. ISPRS- J. Photogramm. Remote. Sens. 186, 218–231.
Xiao, Y., Su, X., Yuan, Q., Liu, D., Shen, H., Zhang, L., 2021. Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection. IEEE Trans. Geosci. Remote Sens. 60, 1–19.
Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.
Xuan, S., Li, S., Han, M., Wan, X., Xia, G.-S., 2019. Object tracking in satellite videos by improved correlation filters with motion estimations. IEEE Trans. Geosci. Remote Sens. 58 (2), 1074–1086.
Yang, J., Guo, J., Yue, H., Liu, Z., Hu, H., Li, K., 2019. CDnet: CNN-based cloud detection for remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 57 (8), 6195–6211.
Yuan, Y., Hu, X., 2015. Bag-of-words and object-based classification for cloud extraction from satellite imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 8 (8), 4197–4205.
Yuan, Y., Wang, Y., Wang, L., Zhao, X., Lu, H., Wang, Y., Su, W., Zhang, L., 2023. Isomer: Isomerous transformer for zero-shot video object segmentation. In: Proc IEEE Int Conf Comput Vision. pp. 966–976.
Zhang, Y., Guindon, B., Cihlar, J., 2002. An image transform to characterize and compensate for spatial variations in thin cloud contamination of Landsat images. Remote Sens. Environ. 82 (2–3), 173–187.
Zhang, J., Wang, H., Wang, Y., Zhou, Q., Li, Y., 2021. Deep network based on up and down blocks using wavelet transform and successive multi-scale spatial attention for cloud detection. Remote Sens. Environ. 261, 112483.
Zhou, T., Li, J., Wang, S., Tao, R., Shen, J., 2020. Matnet: Motion-attentive transition network for zero-shot video object segmentation. IEEE Trans. Image Process. 29, 8326–8338.
Zhou, T., Porikli, F., Crandall, D.J., Van Gool, L., Wang, W., 2022. A survey on deep learning technique for video segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 45 (6), 7099–7122.
Zhu, Z., Woodcock, C.E., 2012. Object-based cloud and cloud shadow detection in Landsat imagery. Remote Sens. Environ. 118, 83–94.
Zhu, Z., Woodcock, C.E., 2014. Automated cloud, cloud shadow, and snow detection in multitemporal landsat data: An algorithm designed specifically for monitoring land cover change. Remote Sens. Environ. 152, 217–234.