International Journal of Geo-Information

MDPI

Article

Channel Attention-Based Multi-Domain Feature Alignment for Moving Vehicle Detection in Satellite Videos Toward Smart Urban Planning

Ning Zhao [1], Xiao Wang [2,3], Xiaopeng Zhang [1], Jun Shi [4], Zhiguo Jiang [2] and Haopeng Zhang [1,2,5,*]

1 School of Astronautics, Beihang University, Beijing 102206, China
2 Tianmushan Laboratory, Beihang University, Hangzhou 311115, China
3 Hangzhou International Innovation Institute, Beihang University, Hangzhou 311115, China
4 School of Software, Hefei University of Technology, Hefei 230601, China
5 Key Laboratory of Spacecraft Design Optimization and Dynamic Simulation Technology, Ministry of Education, Beijing 102206, China
* Correspondence: zhanghaopeng@buaa.edu.cn; Tel.: +86-10-6171-6978

Abstract

Rapid global urbanization is increasing the need for accurate, large-scale traffic monitoring to support sustainable transportation and city governance. Satellite video remote sensing offers a unique way to continuously observe urban road networks over large areas. It provides high-resolution spatio-temporal data that is essential for traffic flow analysis, infrastructure assessment, and dynamic urban planning. Moving vehicle detection in satellite video sequences is a basic task that turns raw imagery into useful traffic-state information, supporting these applications. Despite the advantages of satellite video data, detecting moving vehicles in practice remains a tough problem. Objects are extremely small and lack clear appearance details, while low local contrast makes them hard to separate from complex backgrounds. Satellite platform motion also introduces background misalignment and intensity fluctuations, resulting in missed detections and false alarms that hurt monitoring reliability. Furthermore, current methods do not fully exploit temporal motion cues or transform-domain priors, creating a performance bottleneck that restricts their practical use. To solve these problems, this paper proposes a Channel-Attentive Spatio-Temporal-Frequency Alignment (CASTFA) framework to effectively use and combine multi-dimensional features for moving vehicle detection in satellite videos, with the goal of providing high-quality traffic monitoring data to help smart city planning. Specifically, a State Space-Guided Temporal Compression (SSGTC) module first collects information along the time dimension with linear computational complexity, greatly reducing overhead while keeping motion cues that are critical for traffic-state estimation. The compressed temporal features are then processed with a multi-scale Haar wavelet transform to get hierarchical time-frequency representations that capture subtle motion dynamics across different frequency bands. At the same time, a pre-trained backbone network extracts multi-scale spatial features. To allow these different domains to work together, a Cross-Domain Feature Alignment (CDFA) mechanism aligns and combines spatial and time-frequency features through channel-attentive operations. Experimental results on the publicly available satellite video moving vehicle detection dataset show that the proposed CASTFA method consistently outperforms existing approaches, with better precision, recall, and F1-scores across diverse urban scenarios. These results show that CASTFA can provide reliable moving vehicle detection performance under difficult real-world conditions, supporting accurate traffic-flow monitoring and providing valuable geospatial intelligence for smart urban planning, transportation management, and sustainable city development.

Check for updates

Academic Editors: Wolfgang Kainz, Guoqiang Shen, Qiuxiao Chen, Yanguang Chen, Long Zhou, Yu Liu and Xindong He

Received: 20 May 2026
Revised: 7 July 2026
Accepted: 23 July 2026
Published: 26 July 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Keywords: moving object detection; remote sensing; satellite video; deep learning; smart city; urban planning

1. Introduction

Rapid global urbanization is creating serious challenges for traffic management, infrastructure planning, and sustainable development. Smart urban planning now depends on accurate, large-scale traffic monitoring to guide transportation policy, road network design, and city governance. Traditional systems such as ground-based sensors and surveillance cameras suffer from sparse coverage, high costs, and limited fields of view. It is difficult for them to provide complete traffic-state information across entire urban areas. Satellite video remote sensing offers a promising alternative. It provides a unique way to continuously monitor urban road networks over large areas. By capturing high-resolution, sequential images from space, video satellites supply large-scale spatio-temporal data that is essential for traffic flow analysis, urban mobility assessment, and dynamic transportation planning. Moving vehicle detection in satellite video sequences is a basic task that turns raw imagery into useful traffic-state information and supports these applications.

However, detecting moving vehicles in satellite videos is very different from conventional object detection in natural images, and it faces severe difficulties. As shown in Figure 1, the extremely long imaging distance—typically several hundred kilometers—means that objects of interest often appear as tiny targets that occupy only a few to tens of pixels. This leaves very little useful appearance information such as texture, shape, or structural details. In addition, complex urban scenes contain dense buildings, vegetation, road shadows, and varying illumination. Sensor noise further reduces the intensity and color contrast between moving vehicles and their local backgrounds. As a result, object boundaries become blurred and hard to identify. Single-frame detectors that rely mainly on spatial appearance therefore perform poorly.



Figure 1. Example of moving vehicle detection in VISO dataset.

Beyond these spatial problems, satellite video analysis also faces serious temporal challenges. Although video satellites usually use staring-mode imaging to keep the scene

stable, small platform perturbations, orbital drift, or non-uniform motion can still cause slow background shifts, local geometric distortions, or intensity fluctuations between consecutive frames. These artifacts produce false motion signals that are easily mistaken for real moving targets. Moreover, current remote sensing videos have relatively low frame rates—typically 10 to 25 frames per second for systems such as Jilin-1. This sparse temporal sampling makes motion trajectory modeling and dynamic feature extraction even harder. Because of these combined spatial and temporal difficulties, true moving objects are often hidden by background clutter or confused with motion-like artifacts caused by platform instability. Therefore, robust detection methods must effectively combine spatio-temporal context, use dynamic priors, and suppress platform-induced interference. Such methods are essential for generating high-quality traffic monitoring data to support smart urban planning.

Conventional approaches to moving object detection in satellite video mainly include background modeling [1–3], frame differencing [4–7], and optical flow-based methods [8,9]. Background modeling [1] detects moving objects by estimating a representation of the static scene and identifying deviations from it. Frame differencing [4] locates motion by computing pixel-wise intensity differences between consecutive frames, often followed by thresholding or statistical modeling. Optical flow-based methods [8] estimate per-pixel motion vectors to distinguish foreground motion from the background. These methods use hand-crafted features and fixed assumptions about motion or scene structure. They therefore show limited adaptability to the complex and changing imaging conditions typical of satellite video, such as extreme foreground–background imbalance, illumination changes, platform-induced global motion, and very small object scales. This poor adaptability leads to significant performance bottlenecks and limits their use to coarse motion-based detection without semantic discrimination.

In recent years, object detection in natural images has made significant progress across a variety of tasks [10–16]. These advances offer new ideas for video object detection. However, standard image-based detectors are designed for single-frame analysis and do not use temporal information across video frames. Direct application to satellite videos is therefore inefficient for detecting moving vehicles. To tackle this issue, certain studies [17,18] have combined deep learning with spatio-temporal information. In their framework, motion-related cues are first obtained through conventional methods such as background subtraction or frame differencing, and then fed into deep neural networks to enhance detection accuracy and suppress false alarms. However, these hybrid approaches still depend heavily on traditional preprocessing and do not fully use the feature representation power of deep networks. It is therefore difficult for them to effectively capture motion cues. Some methods [19–22] enhance spatio-temporal feature extraction by constructing auxiliary networks alongside a feature pyramid structure. These approaches typically introduce auxiliary branches next to the main backbone to explicitly model inter-frame motion cues or temporal context dependencies. Although these designs capture rich spatio-temporal dynamics, they often face significant limitations. A major issue is their heavy reliance on 3D convolutions to preserve temporal details. While 3D convolutions are effective for motion modeling, they lead to high computational costs and limit real-time application. In addition, these methods usually work only in the spatial–temporal domain and ignore useful priors from transform domains.

To solve these problems, this paper proposes the Channel-Attentive Spatio-Temporal-Frequency Alignment (CASTFA) framework for moving vehicle detection in satellite videos. The goal is to provide reliable traffic monitoring data to guide smart urban planning and development. Unlike conventional approaches that either use heavy 3D convolutions or ignore transform-domain cues, CASTFA explicitly uses multi-scale Haar wavelet decom-

position to capture discriminative frequency characteristics of motion patterns. At the same time, it keeps computational overhead low through temporal compression. The main contributions are as follows:

1. **State Space-Guided Temporal Compression (SSGTC)**: This module uses state space modeling to compress redundant frames along the temporal dimension. It learns compact latent dynamic representations and efficiently captures long-range motion dependencies with linear computational complexity. This greatly reduces computational cost while keeping subtle motion cues that are essential for detecting small moving objects in complex scenes.
2. **Multi-Scale Frequency Representation (MSFR)**: Instead of working only in the spatial–temporal domain, this module applies Haar wavelet transforms to compressed features. It extracts hierarchical frequency-domain signatures that effectively highlight subtle motion dynamics often lost in low-resolution satellite videos.
3. **Cross-Domain Feature Alignment (CDFA)**: This is a channel-attentive mechanism that adaptively aligns and fuses spatial features from a pre-trained backbone with the derived frequency representations. It enables robust interaction between appearance cues and spectral motion cues.

We evaluate CASTFA on publicly available satellite video datasets for moving vehicle detection. Experimental results show that our method consistently outperforms state-of-the-art approaches, especially in scenarios with tiny objects and noisy motion patterns. These results confirm that explicit frequency-domain modeling is effective for satellite video analysis. The remainder of this paper is organized as follows. Section 2 reviews related work on moving object detection in satellite videos and multi-domain feature modeling. Section 3 details the proposed CASTFA framework, including the SSGTC module, the MSFR module, and the CDFA mechanism. Section 4 presents experimental results and ablation studies on public satellite video datasets to evaluate each key component. Section 5 discusses the practical implications of multi-domain modeling, analyzing the trade-offs between detection accuracy and computational efficiency and the inherent limitations of the current architecture. Finally, Section 6 summarizes the contributions of this work.

2. Related Works

2.1. Traditional Object Detection Method

In moving object detection research, pixels belonging to moving objects are typically defined as foreground, while those corresponding to static objects and irrelevant dynamic elements are classified as background. To separate foreground regions from video sequences, background subtraction has become a classic and widely adopted technique. This approach builds a background model for each pixel, characterizing the typical range of its intensity values in the absence of moving objects. When a new frame arrives, a pixel is labeled as foreground if its observed value deviates significantly from this model. The simplest background models usually estimate the background using the average or median of multiple video frames [23]. To address these limitations, more sophisticated methods have been developed. Zhang et al. [1] proposed an Extended Low-Rank and Structured Sparse Decomposition (E-LSD) method that incorporates bounded error modeling into foreground-background matrix decomposition, explicitly accounting for unexplained residuals in low-resolution satellite videos to enhance moving object detection accuracy. Zhang et al. [2] proposed an online low-rank and structured sparse decomposition (O-LSD) method that enables real-time moving object detection in satellite videos by reformulating batch optimization into a frame-wise stochastic process, significantly reducing processing delay while maintaining accuracy.

Alongside background modeling, frame differencing represents a classical and widely used approach for motion detection in satellite videos. Rather than constructing an explicit background model, it identifies moving regions by computing pixel-wise intensity differences between consecutive or adjacent frames. Owing to its simplicity and low computational cost, frame differencing is often employed as a preliminary step to generate motion cues. Zhang et al. [4] proposed a motion-based object detection method that leverages multi-frame motion continuity to detect and associate targets. By analyzing the motion induced by platform attitude and compensating for it, the background is stabilized; subsequently, an adaptive thresholding strategy—guided by detection results from the previous frame—is employed to segment candidate regions, while trajectory association based on motion prediction enables robust tracking of objects even when they temporarily disappear due to brightness fluctuations. Li et al. [5] proposed a vessel detection and tracking method based on multi-scale saliency and surrounding contrast analysis. By constructing a Dynamic Multi-Scale Saliency Map (DMSM), their approach effectively overcomes the issue of internal holes commonly encountered in traditional frame differencing when dealing with targets exhibiting small displacements, thereby achieving more robust extraction of moving vessel regions. Shi et al. [6] proposed a moving airplane detection method based on Normalized Frame Difference Labeling (NFDL). This method adopts a non-recursive strategy to avoid the dependence of traditional background modeling on the number of initial video frames. It uses the first or last frame of the video as a distant reference frame to effectively mitigate the foreground aperture problem caused by small target displacements. The normalized frame difference image is segmented into three clusters via k-means clustering, enabling simultaneous detection of both bright and dark moving targets. Furthermore, Laplacian second-order derivative boundary analysis is incorporated to distinguish genuine moving objects from static background interference, thereby enhancing detection completeness and robustness.

Optical flow-based methods estimate per-pixel motion vectors between consecutive frames to characterize object movement. By analyzing the magnitude and direction consistency of these flow fields, such techniques can effectively identify coherent moving regions without requiring explicit background modeling. Cao et al. [8] proposed a video object detection method that leverages optical flow–guided sparse feature propagation to efficiently transfer deep features from keyframes to non-keyframes, significantly accelerating inference, while simultaneously employing dense feature aggregation across temporally adjacent frames to enhance detection robustness against motion blur, occlusion, and low resolution.

Traditional moving object detection methods demonstrate certain effectiveness under specific conditions. Optical flow methods can accurately capture pixel-level motion details and are well-suited for tasks requiring high motion continuity. Background subtraction approaches directly extract foreground objects by modeling static backgrounds and perform reliably in scenes with stable backgrounds. Frame differencing, owing to its computational simplicity and high efficiency, is often employed for rapid preliminary detection of motion regions. However, these methods exhibit significant limitations when applied to complex satellite video scenarios. Optical flow is computationally expensive and highly sensitive to illumination changes. Background subtraction is prone to interference from background jitter, occlusion, and abrupt lighting variations, often leading to numerous false alarms. Frame differencing, on the other hand, tends to produce internal holes in detected objects when dealing with slow-moving or minimally displaced targets, compromising structural integrity. Consequently, although these conventional techniques may be adequate for standard video sequences, their robustness and adaptability remain insufficient in the

highly dynamic and interference-rich environment of remote sensing video, calling for more specialized and effective algorithmic designs.

2.2. Object Detection Method Based on Deep Learning

In recent years, object detection in natural images has witnessed remarkable progress, largely driven by the advent and evolution of deep learning. Since deep neural networks demonstrated their superiority in large-scale image recognition tasks, they have become the dominant paradigm for detecting objects in complex, real-world scenes—surpassing traditional methods based on handcrafted features in both accuracy and robustness. Today, deep learning-based approaches firmly hold a leading position in computer vision, underpinning a wide range of applications from autonomous systems to video analysis. Broadly speaking, current object detection frameworks can be categorized into two main families: those that adopt a two-stage strategy [10–12], first generating candidate regions and then refining them for classification and localization; and those that follow a one-stage design [14–16], directly predicting object categories and bounding boxes in a single, efficient pass. This methodological dichotomy reflects the ongoing balance between detection accuracy and computational efficiency, and together these approaches have significantly advanced the state of the art in natural image understanding—providing a solid foundation for specialized tasks such as moving object detection in dynamic environments.

Beyond natural images, deep learning has also been widely applied to aerial and satellite image recognition tasks. In aerial imagery, person detection for search and rescue has attracted considerable attention, with comprehensive surveys and benchmark datasets established to facilitate systematic evaluation [24], and lightweight distillation networks developed to enable efficient edge deployment under resource-constrained conditions [25]. In optical satellite imagery, salient maritime ship detection has been advanced through coarse-to-fine Transformer architectures that address extreme scale variations and background interference via multi-stage attention-driven refinement [26]. While these works tackle the shared challenge of detecting small and visually ambiguous objects in remote sensing scenes, they operate on static imagery and do not exploit inter-frame motion information, which is critical for distinguishing moving targets from cluttered backgrounds in video sequences.

For satellite video sequences, where moving vehicles are extremely tiny, visually ambiguous, and easily confused with complex backgrounds, single-frame appearance information alone is often insufficient for reliable detection. To exploit the complementary temporal dimension, LaLonde et al. [27] proposed a two-stage framework. The method first leverages multi-frame inputs to jointly model spatial appearance and temporal motion consistency inside a convolutional architecture, enabling the network to generate candidate regions. Subsequently, it performs high-resolution analysis within these candidate regions and uses heatmap regression to simultaneously predict the precise center locations of all objects present. Feng et al. [28] build upon the CornerNet [15] architecture by introducing a Cross-Frame Module (CFM) that leverages cross-frame differencing to effectively fuse temporal features from adjacent frames with the appearance features of the current frame. This integration significantly improves recall for small and blurry moving vehicles in satellite videos. Xiao et al. [19] proposed DSFNet, a two-stream deep network for moving object detection in satellite videos that explicitly integrates complementary static and dynamic cues. The method employs a 2D stream to extract contextual appearance information from a single frame and a lightweight 3D stream that efficiently captures spatio-temporal motion patterns across consecutive frames. Chen et al. [21] proposed MICPL, which builds upon DSFNet [19] by introducing motion patterns as a learnable implicit supervisory signal

during training and employs a long short-term memory [29]-based model to capture global spatio-temporal dependencies from multi-frame visual sequences.

Existing deep learning-based methods for moving object detection in satellite videos, while effective to some extent, suffer from two key limitations. First, their cross-domain interaction and fusion strategies are simplistic, typically relying on late-stage feature concatenation or element-wise addition, which fails to model the deep semantic interplay between appearance and motion cues. Second, these approaches place excessive emphasis on temporal-domain features derived directly from raw pixel sequences, while largely ignoring the potential of integrating complementary representations from other transform domains. This narrow focus limits their ability to capture complex motion patterns—such as non-uniform motion, occlusion recovery, or subtle vibrations—especially under low-frame-rate and high-noise conditions typical of satellite imagery, thereby constraining both robustness and generalization.

3. Method

3.1. Overview

Figure 2 illustrates the overall framework of our proposed CASTFA method. The framework first employs three parallel branches to extract multi-scale features from the spatial domain, transform domain, and temporal domain of the current frame and previous frames, respectively. The extracted features are then fed into a channel attention module to facilitate interactive integration of multi-domain information. Subsequently, the multi-domain interactive features at multiple scales are passed to a multi-scale feature integration module to enable cross-scale feature interaction. The resulting features are processed by a lightweight convolutional layer to produce the final regression output and heatmap, which is then used for loss computation.

3.2. Multi-Scale Frequency Representation (MSFR)

Owing to the superior performance of anchor-free detection architectures in remote sensing moving object detection, our network is built upon CenterNet [14] and adopts the DLA-34 [30] pretrained on ImageNet [31] as the backbone for spatial feature extraction. For frequency-domain feature extraction, we employ the Haar wavelet transform to convert spatial-domain features into the frequency domain. As one of the earliest and most fundamental wavelet transforms, 2D Haar wavelet transform holds significant importance in multi-scale signal analysis and image processing. It pioneered the concept of orthogonal wavelet decomposition, enabling explicit separation of image content into low-frequency approximations that capture global structure and high-frequency details that encode directional edges and local variations. Owing to its minimal computational complexity, reliance solely on addition, subtraction, and bit-shift operations, and inherent differentiability, the 2D Haar wavelet transform has remained widely adopted across decades.

To address the limitation of conventional convolutional operations in capturing directional characteristics of objects in remote sensing imagery, we incorporate 2D Haar wavelet transform as a complementary frequency-domain analysis mechanism within our detection framework, whose structure is depicted in Figure 3. The Haar wavelet, as the simplest member of the Daubechies wavelet family, provides an efficient way to decompose images into distinct frequency subbands that explicitly encode scale, texture, and angular information. Its mother wavelet function $\psi(t)$ is defined as a piecewise constant function:

$$\psi(t) = \begin{cases} 1, & 0 \le t < \frac{1}{2} \\ -1, & \frac{1}{2} \le t < 1 \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

Correspondingly, the scaling function $\phi(t)$ and the associated filter coefficients $h[n]$ are expressed as:

$$\phi(t) = \begin{cases} 1, & 0 \le t < 1 \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

$$h[n] = \begin{cases} \frac{1}{\sqrt{2}}, & \text{if } n = 0, 1 \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

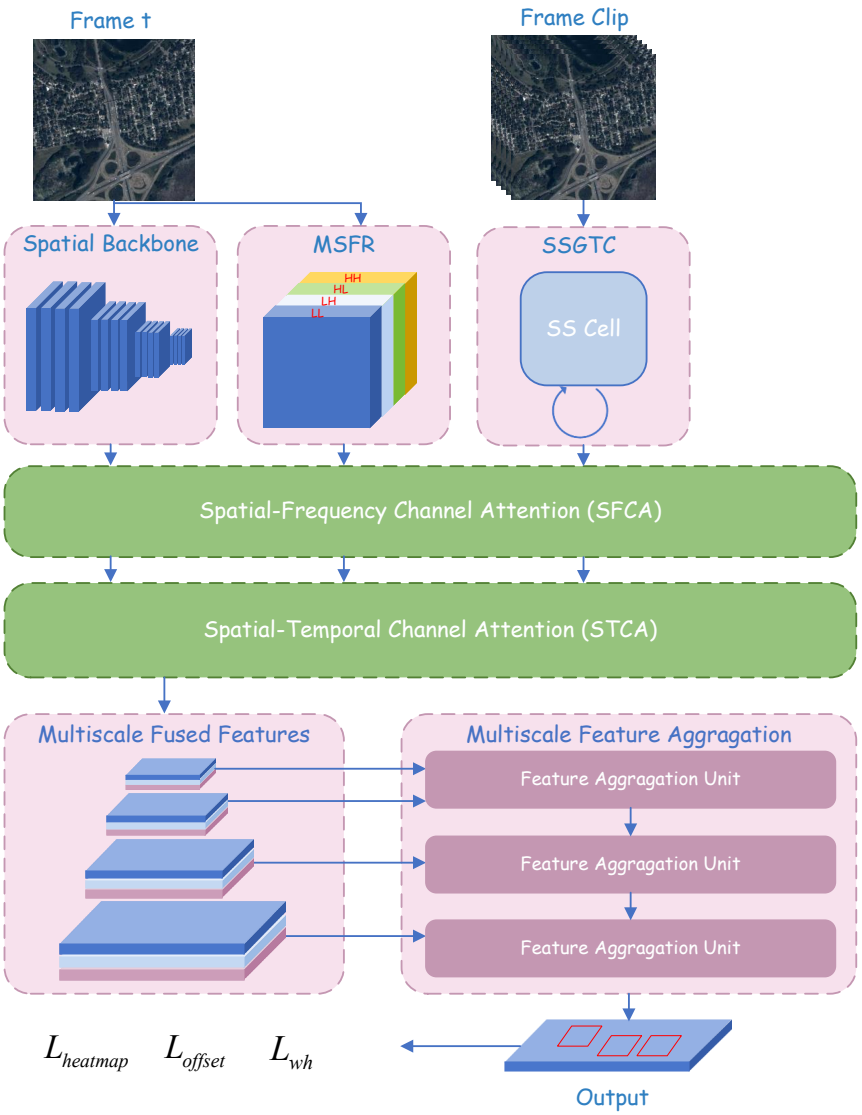


Figure 2. The overall architecture of the proposed method.

Using these filter coefficients, the mother wavelet function can be represented as a linear combination of shifted scaling functions:

$$\frac{1}{\sqrt{2}}\psi\left(\frac{t}{2}\right) = \frac{1}{\sqrt{2}}(\phi(t-1) - \phi(t)) \tag{4}$$

For a 2D image block $\mathbf{A}$ of size $N \times N$, the discrete Haar wavelet transform is mathematically formulated as $\mathbf{B} = \mathbf{HAH}^T$, where $\mathbf{H}$ denotes the N-order discrete Haar transform matrix. The inverse transform, which reconstructs the original spatial-domain representation, is given by $\mathbf{A} = \mathbf{H}^T\mathbf{BH}$. To construct the transform matrix $\mathbf{H}$, we define an integer

parameter $k = 2^p + q - 1$ with $0 \le p \le n-1$, where $q = 0$ or 1 when $p = 0$, and $1 \le q \le 2^p$ when $p \neq 0$. Based on this parameterization, the Haar basis functions are expressed as:

$$h_{0z} = h_{00z} = \frac{1}{\sqrt{N}}, \quad z \in [0,1] \tag{5}$$

$$h_k(z) = h_{pq}(z) = \frac{1}{\sqrt{N}} \begin{cases} 2^{\frac{p}{2}}, & \frac{q-1}{2^p} \le z < \frac{q-0.5}{2^p} \\ -2^{\frac{p}{2}}, & \frac{q-0.5}{2^p} \le z < \frac{q}{2^p} \\ 0, & \text{otherwise in } [0,1] \end{cases} \tag{6}$$

Each level of the Haar wavelet decomposition splits the input into four distinct subbands: a low-frequency subband (LL) preserving the overall structural information, and three high-frequency subbands—LH capturing vertical edges, HL capturing horizontal edges, and HH capturing diagonal textures. Subsequent decomposition levels operate recursively only on the LL subband, enabling multi-level analysis. A critical property of this hierarchical decomposition is that the spatial resolution of the feature maps is halved after each transformation level. This natural downsampling behavior aligns precisely with the progressive resolution reduction observed in deep convolutional neural networks, where feature maps become coarser as network depth increases. This inherent compatibility between wavelet-based frequency decomposition and CNN feature hierarchy forms the foundation for seamless multiscale spatial-frequency feature fusion in our detection framework.

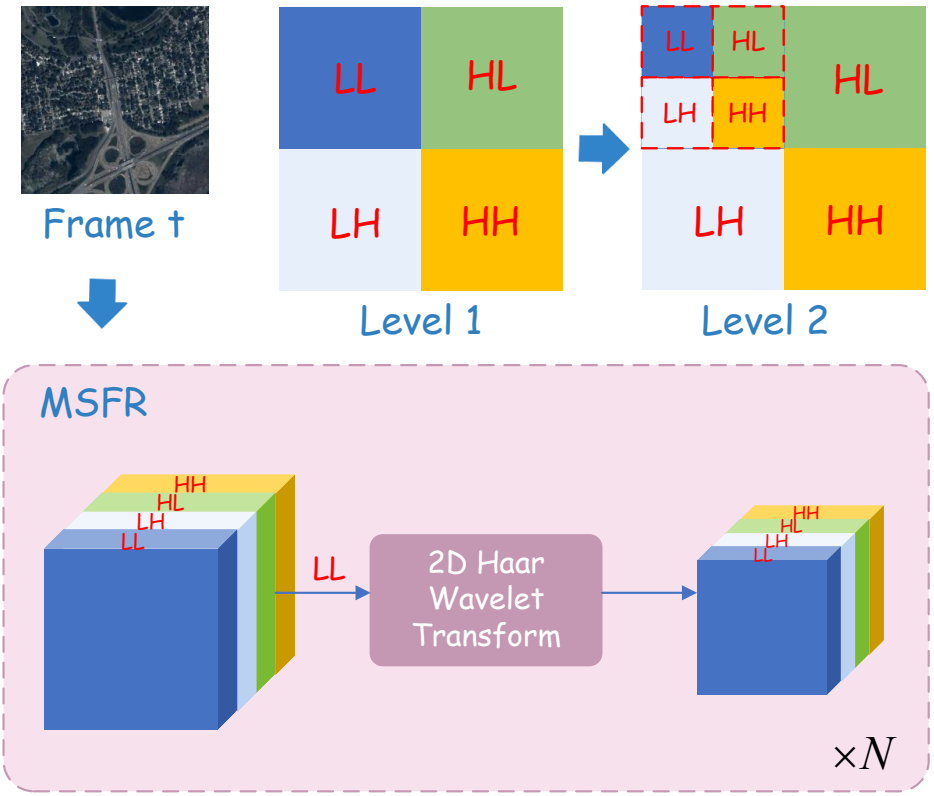


Figure 3. The structure of MSFR module.

3.3. State Space-Guided Temporal Compression (SSGTC)

In remote sensing video object detection, the discrimination of moving targets relies not only on inter-frame displacement cues but, more fundamentally, on the quality of spatial feature representation within individual frames. Due to the typically small scale, low contrast, and complex background interference prevalent in remote sensing imagery,

insufficient spatial feature extraction often leads to ambiguous target-background boundaries, which in turn undermines the reliability of subsequent temporal analysis—even when sophisticated temporal modeling is applied. Conventional convolutional operations, constrained by their inherently local receptive fields, struggle to capture long-range spatial dependencies between target regions and distant contextual structures, resulting in features with limited discriminative power.

Therefore, a new temporal modeling module that leverages the selective state space mechanism to efficiently capture global spatial correlations with linear computational complexity is integrated into our framework. The core of this module is built upon the continuous-time state space model, which characterizes the evolution of latent states through a linear differential equation system:

$$\frac{d}{dt}x(t) = \mathbf{A}x(t) + \mathbf{B}u(t), \quad y(t) = \mathbf{C}x(t) + \mathbf{D}u(t) \tag{7}$$

where $x(t) \in \mathbb{R}^N$ represents the hidden state, $u(t) \in \mathbb{R}$ is the input signal, $y(t) \in \mathbb{R}$ is the output, and $\mathbf{A} \in \mathbb{R}^{N\times N}$, $\mathbf{B} \in \mathbb{R}^{N\times 1}$, $\mathbf{C} \in \mathbb{R}^{1\times N}$, $\mathbf{D} \in \mathbb{R}$ are learnable parameters governing state transitions and input-output mappings. To apply this continuous model to discrete visual features, the zero-order hold discretization with step size Δ is adopted to adapt the continuous model to the discrete space:

$$x_k = \bar{\mathbf{A}}x_{k-1} + \bar{\mathbf{B}}u_k, \quad y_k = \mathbf{C}x_k + \mathbf{D}u_k \tag{8}$$

where $\bar{\mathbf{A}} = e^{\mathbf{A}\Delta}$ and $\bar{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(e^{\mathbf{A}\Delta} - \mathbf{I})\mathbf{B}$. Unlike conventional state space models with fixed parameters, our module incorporates the selective mechanism proposed in Mamba [32], wherein the system parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, and the discretization step Δ are dynamically generated as functions of the input feature. This selectivity enables the model to adaptively emphasize spatial regions with strong semantic relevance while suppressing background clutter or irrelevant textures during state propagation. When processing 2D feature maps, we first flatten spatial dimensions into a sequential representation and perform bidirectional selective scanning along horizontal and vertical directions. This design allows the model to establish global receptive fields across the entire feature map with only $\mathcal{O}(L)$ computational complexity, where L denotes the spatial sequence length, in contrast to the $\mathcal{O}(L^2)$ complexity of self-attention mechanisms. By constructing such long-range spatial dependencies within individual frames, our module produces structurally coherent feature representations that preserve both local details and global context, thereby providing a solid foundation for subsequent temporal aggregation across video frames.

The details of our SSGTC module are illustrated in Figure 4, which adopts a streamlined recurrent architecture that integrates selective state space modeling with gated temporal dynamics. Starting from the conventional ConvLSTM [33] formulation, we eliminate all learnable convolutional kernels and bias terms within the gating mechanisms to reduce parameter redundancy while preserving the core temporal propagation structure. This simplification yields a parameter-free gating operation where input, forget, and output gates share identical activation values:

$$i_t = f_t = o_t = \sigma(\mathbf{X}_t + \mathbf{H}_{t-1}) \tag{9}$$

$$\mathbf{C}_t = f_t \odot \mathbf{C}_{t-1} + i_t \odot \tanh(\mathbf{X}_t + \mathbf{H}_{t-1}) \tag{10}$$

$$\mathbf{H}_t = o_t \odot \tanh(\mathbf{C}_t) \tag{11}$$

As illustrated in Figure 4, SSGTC processes each frame through two complementary pathways: (1) a horizontal propagation path that maintains temporal continuity via

cell state $\mathbf{C}_t$ updates, capturing motion-evolved evidence across frames; (2) a vertical spatial modeling path where visual state space blocks (VSB) extract long-range spatial dependencies across individual frames. This orthogonal design decouples spatial context modeling from temporal evolution, preventing interference between the two distinct modeling objectives.

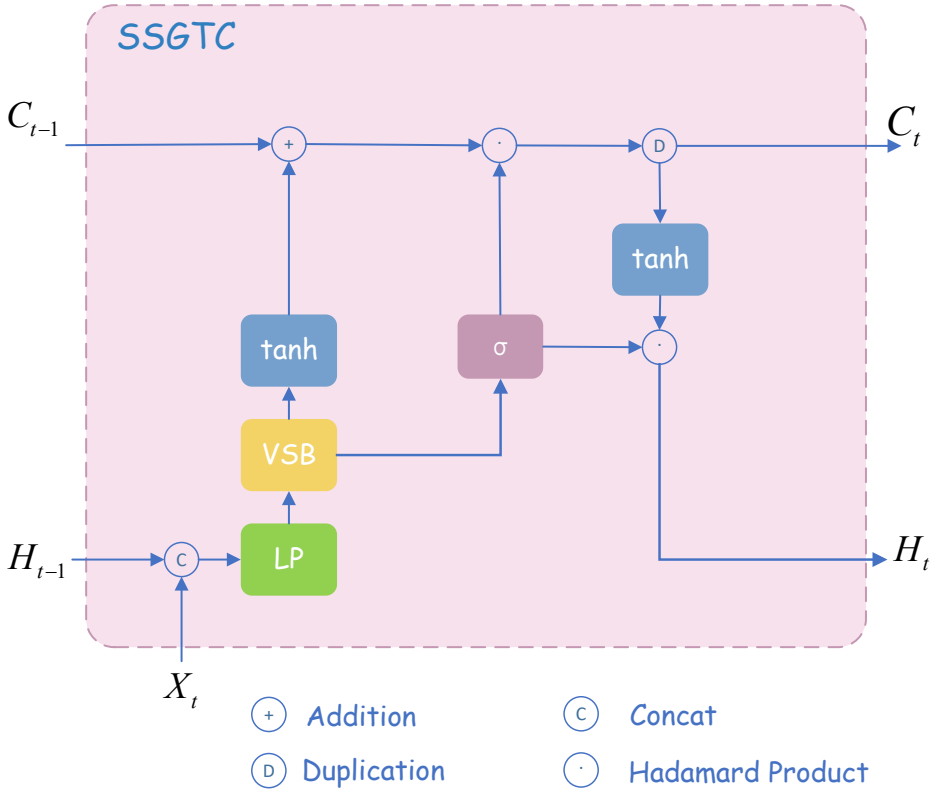


Figure 4. The structure of SSGTC module.

The spatial enhancement pathway begins by fusing the current input feature $\mathbf{X}_t$ with the previous hidden state $\mathbf{H}_{t-1}$ through a linear projection layer, producing a unified representation that encodes both current appearance and historical context. This fused representation is then processed by the VSB to establish global spatial correlations with linear complexity. The resulting spatially enriched features subsequently drive the gating and state update operations:

$$\mathbf{F}_t = \sigma(\text{VSB}(\text{LP}(\mathbf{X}_t, \mathbf{H}_{t-1}))) \tag{12}$$

$$\mathbf{C}_t = \mathbf{F}_t \odot (\tanh(\text{VSB}(\text{LP}(\mathbf{X}_t, \mathbf{H}_{t-1}))) + \mathbf{C}_{t-1}) \tag{13}$$

$$\mathbf{H}_t = \mathbf{F}_t \odot \tanh(\mathbf{C}_t) \tag{14}$$

where $\text{LP}(\cdot)$ denotes the linear projection operation.

The VSB adopts a dual-stream parallel architecture to efficiently model global spatial dependencies within feature maps. As illustrated in Figure 5, the input features first pass through an initial linear embedding layer and are subsequently split into two distinct information pathways. The primary stream traverses a 3 × 3 depth-wise convolution layer followed by a SiLU activation function to capture local contextual cues, and then proceeds into the two-dimensional selective scan module (SS2D) for global relationship modeling. The secondary stream serves as a residual branch that bypasses the SS2D module and undergoes direct SiLU activation. After the SS2D processing, the primary stream is

refined through layer normalization and subsequently fused with the residual branch via element-wise addition to produce the final block output.

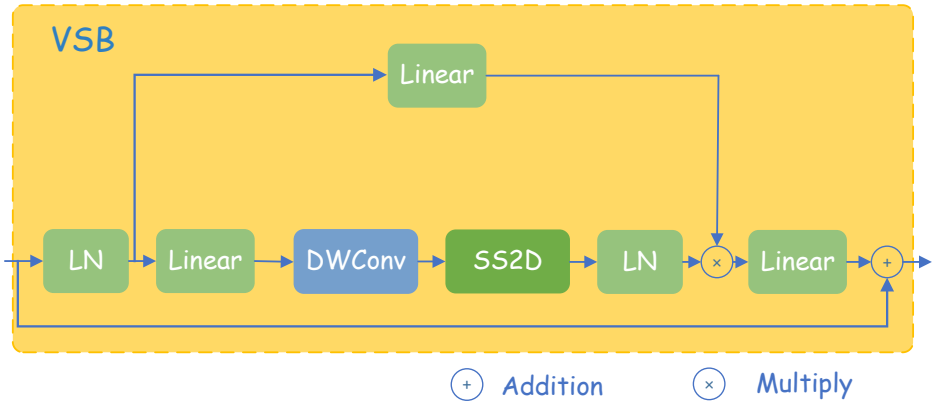


Figure 5. The structure of VSB.

This architectural design diverges from the conventional vision transformer paradigm, which typically follows a "normalization → self-attention → normalization → MLP" sequence within each block. By deliberately omitting the computationally intensive MLP stage, the VSB achieves significantly reduced complexity while preserving spatial modeling capacity. This simplification enables the stacking of a greater number of blocks under comparable model depth constraints, thereby enhancing hierarchical feature representation.

To accommodate two-dimensional image structures, the VSB first reshapes the serialized feature representation $\mathbf{z} \in \mathbb{R}^{B\times L\times C}$ back to its spatial layout $\mathbb{R}^{B\times H\times W\times C}$. Subsequently, the SS2D mechanism establishes long-range spatial dependencies through multi-directional scanning. Specifically, the feature map is unfolded along four orthogonal directions: top-left to bottom-right, bottom-right to top-left, top-right to bottom-left, and bottom-left to top-right, generating four distinct sequential representations:

$$\mathbf{z}_v = \text{expand}(\mathbf{z}, v), \quad v \in \{1, 2, 3, 4\} \tag{15}$$

Each directional sequence is independently processed by the selective state space operator (S6), which dynamically generates system parameters based on input content to adaptively focus on semantically relevant regions:

$$\bar{\mathbf{z}}_v = \text{S6}(\mathbf{z}_v) \tag{16}$$

The processed sequences are then reconstructed into their original two-dimensional spatial configuration through a merge operation:

$$\bar{\mathbf{z}} = \text{merge}(\bar{\mathbf{z}}_1, \bar{\mathbf{z}}_2, \bar{\mathbf{z}}_3, \bar{\mathbf{z}}_4) \tag{17}$$

Multiple SSGTC modules are stacked and PatchMerge operations are inserted between them for downsampling, thereby constructing multi-scale temporal features that can be fused with other branches.

3.4. Channel Attention-Based Cross-Domain Feature Alignment

To comprehensively characterize the representation of moving objects, multi-dimensional features are extracted from the spatial, frequency, and temporal domains. The spatial domain captures geometric structures and appearance details, the frequency domain reveals spectral distribution patterns of textures and motion responses, and the temporal-domain models the dynamic evolution of inter-frame motion. Since the discriminative cues

embedded in these distinct domains exhibit strong complementarity, simply concatenating or independently utilizing them fails to fully exploit their synergistic potential. An effective fusion mechanism is therefore essential to integrate multi-domain features coherently, thereby enhancing both the robustness and accuracy of motion detection. The following subsection details the proposed fusion strategy.

We adopt a progressive fusion strategy to integrate multi-domain features. Taking the features at the i-th layer as an example, we first employ a convolutional channel attention module, termed Spatial-Frequency Channel Attention (SFCA), to separately fuse the frequency-domain feature with the spatial-domain and temporal-domain features. As illustrated in Figure 6, the SFCA module begins by reshaping the frequency-domain feature through a bottleneck structure to align its channel dimension with that of the target-domain feature. Subsequently, a squeeze-and-excitation (SE) block [34] is applied to the expanded feature for channel-wise attention weighting. The weighted feature is then added to the corresponding target domain feature (spatial or temporal), followed by an additional SE block to further refine the fused representation. This process yields two intermediate fused features: the spatial-frequency fused feature $\mathbf{F}_i^{s/f}$ and the temporal-frequency fused feature $\mathbf{F}_i^{t/f}$.

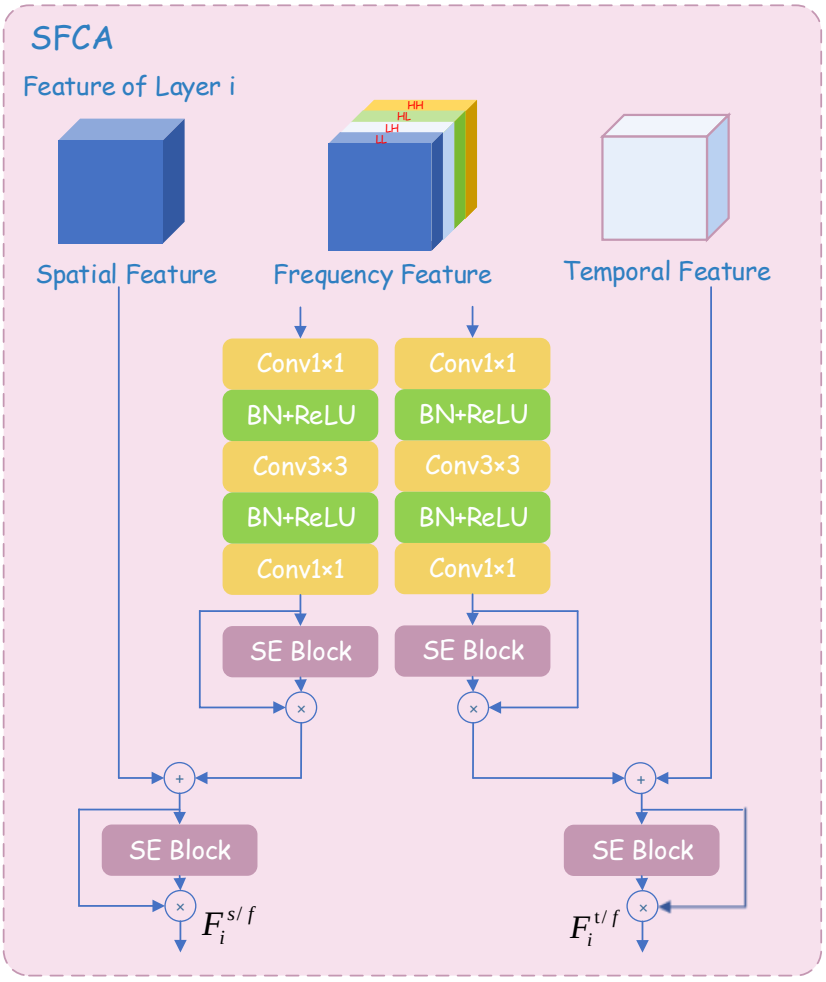


Figure 6. The structure of SFCA.

The SE block implements channel attention through global context modeling and adaptive reweighting. Given an input feature map $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$, the channel-wise statistics are first aggregated via global average pooling:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j), \quad c = 1,2,\ldots,C, \tag{18}$$

where z_c denotes the aggregated value for the c-th channel. The excitation operation then learns channel dependencies through two fully connected layers with a reduction ratio r:

$$\hat{\mathbf{z}} = \sigma\big(\mathbf{W}_2\, \delta(\mathbf{W}_1 \mathbf{z})\big), \tag{19}$$

where $\mathbf{z} = [z_1, z_2, \ldots, z_C]^\top$, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ are the weight matrices, $\delta(\cdot)$ denotes the ReLU activation, and $\sigma(\cdot)$ represents the sigmoid function. The final output is obtained by rescaling the original feature channels:

$$\tilde{\mathbf{x}}_c = \hat{z}_c \cdot \mathbf{x}_c, \quad c = 1,2,\ldots,C. \tag{20}$$

The SE block essentially functions as a single-stream self-attention mechanism that performs channel-wise recalibration solely based on the feature's own global statistics, lacking the capability to explicitly model interdependencies between distinct feature streams. In contrast, the preliminarily fused features $\mathbf{F}_i^{s/f}$ and $\mathbf{F}_i^{t/f}$ already encode complementary cues from spatial-frequency and temporal-frequency perspectives, respectively. Their interaction should therefore transcend isolated intra-stream channel enhancement and establish explicit cross-domain correlations. To this end, we introduce a Spatial–Temporal Channel Attention (STCA) module, as illustrated in Figure 7, which implements channel cross-attention by treating features from one domain as Queries and those from the other as Keys and Values.

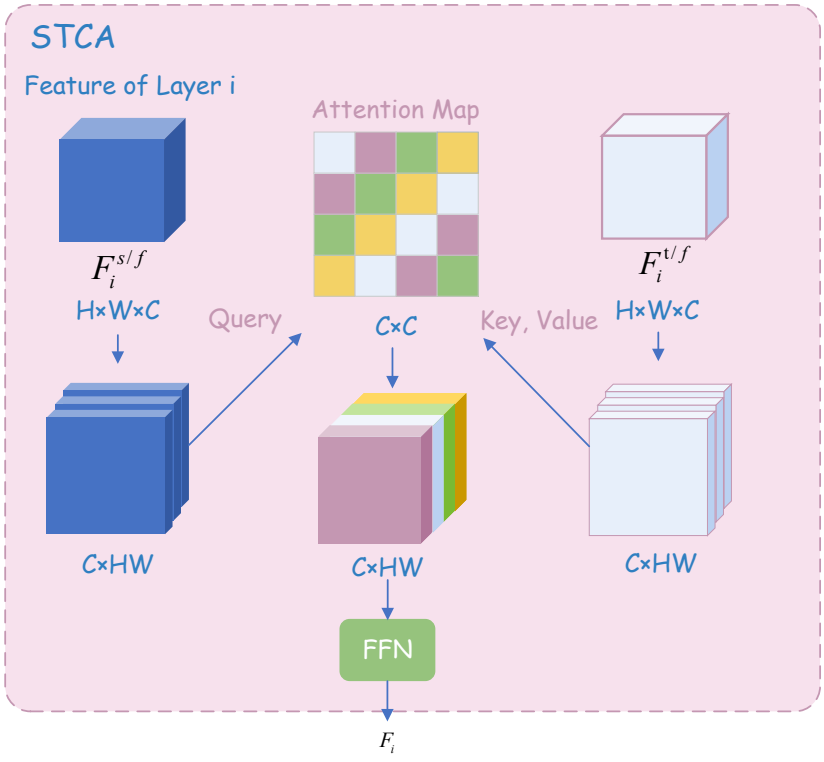


Figure 7. The structure of STCA.

In detail, channel-wise projections are first applied to both feature streams:

$$\mathbf{Q} = \mathrm{Conv}^{q}_{1\times1}(\mathbf{F}_i^{s/f}), \quad \mathbf{K} = \mathrm{Conv}^{k}_{1\times1}(\mathbf{F}_i^{t/f}), \quad \mathbf{V} = \mathrm{Conv}^{v}_{1\times1}(\mathbf{F}_i^{t/f}), \tag{21}$$

where each convolution shares parameters spatially and performs a linear transformation $\mathbb{R}^{C\to C/r}$ solely along the channel dimension.

Cross-domain attention weights are computed and used to aggregate information:

$$\tilde{\mathbf{z}} = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{C/r}}\right)\mathbf{V}, \tag{22}$$

followed by a feed-forward network to enhance the representation. The refined attention map then recalibrates the target feature $\mathbf{F}^{t/f}$ via channel-wise broadcasting multiplication:

$$\mathbf{F}_i = \mathrm{FFN}(\tilde{\mathbf{z}}) \tag{23}$$

This design enables bidirectional, content-aware feature selection and explicit cross-domain interaction between dual-domain features, ultimately yielding a more discriminative joint representation $\mathbf{F}_i$.

4. Experimental Results and Analysis

4.1. Dataset

To validate the effectiveness of our method, we conduct experiments on the publicly available **RSCarData** dataset for remote sensing video object detection.

The **RSCarData** dataset was initially proposed by Xiao et al. [19]. It adopts the test set from the VISO [35] dataset as its test set, comprising 7 video sequences of size 1024 × 1024. For the training set, the authors collected additional remote sensing video data from the Jilin-1 satellite, consisting of 72 video sequences cropped to 512 × 512 resolution. Each image is annotated with horizontal bounding boxes. Both the training and test sets have a frame rate of 10 fps. The dataset contains only a single class of moving vehicles, with extremely small target scales—averaging approximately 40 pixels in area—and limited appearance features, posing significant challenges for object detection tasks.

To further validate the cross-dataset generalization capability of CASTFA, we additionally employ the car subset of the **SAT-MTB** dataset [36] as an independent benchmark. SAT-MTB is currently the largest publicly available satellite video dataset with multitask annotations, comprising 249 video sequences and over 50,000 frames captured by the Jilin-1 03 satellites across 14 countries and regions worldwide. The spatial resolution is 0.92 m at the sub-satellite point, and the frame rate is 10 fps. For moving vehicle detection, we utilize the tracking-task annotations of the car category, which provide frame-by-frame horizontal bounding boxes for moving cars. Compared to RSCarData, SAT-MTB presents a more challenging testbed due to its more stringent imaging conditions—including lower object contrast, smaller object scales, and greater scene diversity spanning multiple continents and climate zones—thus providing a rigorous assessment of model generalizability.

4.2. Evaluation Metrics

To quantitatively evaluate the performance of moving object detection in satellite videos, we adopt three standard metrics: precision, recall, and F1 score. These metrics are computed based on the intersection over union (IoU) between predicted bounding boxes and ground-truth annotations. Specifically, a detection is considered a true positive (TP) if its IoU with any ground-truth box exceeds a predefined threshold τ_{IoU}; otherwise, it is

counted as a false positive (FP). Ground-truth objects without any matching detection are counted as false negatives (FN). The metrics are defined as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{24}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{25}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{26}$$

where TP denotes correctly detected moving vehicles, FP represents false alarms (e.g., misdetected background regions or motion artifacts), and FN indicates missed detections.

Due to the extremely small scale of moving vehicles in satellite videos, minor deviations in bounding box localization or center point offsets can lead to significant IoU degradation. To balance localization accuracy and practical detection requirements for tiny moving objects, we set the IoU threshold to $\tau_{\text{IoU}} = 0.3$, which adequately reflects both target localization capability and bounding box fitting quality in this challenging scenario.

4.3. Experiment Setup

In our experiments, we employ a pre-trained DLA-34 [30] to extract multi-scale spatial-domain features. Given an input image with spatial resolution $H \times W$ and a base channel dimension C (where $C = 16$ in our implementation), the network first processes the input through a base convolutional layer, yielding an initial feature map of spatial size $\frac{H}{2} \times \frac{W}{2}$ with C channels. Subsequently, the architecture progresses through five hierarchical levels (level 1 to level 5), where each stage performs a 2× spatial downsampling while simultaneously doubling the channel dimension, thereby constructing a coarse-to-fine multi-scale feature pyramid. The dimensional specifications of feature maps at each stage are as follows: the base layer outputs $\frac{H}{2} \times \frac{W}{2} \times C$, level 1 produces $\frac{H}{4} \times \frac{W}{4} \times 2C$, level 2 yields $\frac{H}{8} \times \frac{W}{8} \times 4C$, level 3 generates $\frac{H}{16} \times \frac{W}{16} \times 8C$, level 4 outputs $\frac{H}{32} \times \frac{W}{32} \times 16C$, and level 5 produces $\frac{H}{64} \times \frac{W}{64} \times 32C$. This hierarchical design enables the network to simultaneously capture global contextual information for large-scale objects and fine-grained local details for small-scale objects, providing rich multi-scale representations essential for moving object detection tasks.

For the MSFR branch, we adopt a three-stage downsampling architecture. After each wavelet transform operation, the LL subband is selected as the input for the subsequent stage. The four subbands—LL, HL, LH, and HH—generated at each stage are concatenated along the channel dimension to form a multi-frequency representation, which is then utilized for cross-domain feature fusion.

For the SSGTC branch, we similarly construct multi-scale outputs and hidden states through a stack of SSGTC modules. The output of each SSGTC module undergoes a PatchMerge operation, which performs information-preserving downsampling without feature loss, and the resulting representation serves as the input to the subsequent stage. The base embedding dimension is set to 96.

After completing feature fusion across spatial, frequency, and temporal domains, adjacent-scale features are hierarchically integrated through a feature fusion block. Denoting the high-level feature as F_{l+1} and the low-level feature as F_l, the fusion process is concisely formulated as:

$$P_l = \mathcal{D}_3\big(\mathcal{U}(\mathcal{D}_1(F_{l+1})) \oplus \mathcal{D}_2(F_l)\big) \tag{27}$$

where $\mathcal{D}_1$ and $\mathcal{D}_2$ perform deformable convolutions on the high-level and low-level features respectively to achieve adaptive spatial alignment, $\mathcal{U}$ denotes upsampling, $\oplus$ represents element-wise addition, and $\mathcal{D}_3$ refines the fused result through nonlinear transformation. Specifically, the high-level semantic feature first undergoes deformable convolution for geometric rectification, followed by upsampling to match the spatial resolution of the low-level feature. Concurrently, the low-level detail feature is adaptively refined via deformable convolution. The two spatially aligned features are then fused through element-wise addition, and the fused representation is further enhanced by a third deformable convolution to produce the multi-scale feature P_l, which jointly preserves high-level discriminability and low-level spatial precision. This dual-stage deformable mechanism—combining pre-fusion rectification and post-fusion refinement—effectively mitigates spatial misalignment in cross-scale fusion, thereby significantly improving localization accuracy and robustness for moving object detection.

The detection head employs three independent prediction streams operating concurrently to estimate the class-specific heatmap, center point offsets, and bounding box dimensions. Each stream shares an identical architecture comprising a 3 × 3 convolutional layer with 128 output channels, followed by a ReLU activation and a final 1 × 1 convolutional layer that produces the task-specific output. Detection instances are ultimately generated by decoding and integrating the predictions from these three streams into spatially localized bounding boxes.

The loss function design employs a weighted combination of heatmap loss, size regression loss, and offset loss to supervise end-to-end training, with each component jointly optimizing semantic confidence, geometric dimensions, and spatial precision of object localization. The heatmap loss adopts a modified Focal Loss formulation to mitigate the extreme foreground-background imbalance:

$$\mathcal{L}_{hm} = \frac{1}{N} \sum_{x,y,c} \begin{cases} (1 - \hat{p}_{c,x,y})^{\alpha} \, |p_{c,x,y} - \hat{p}_{c,x,y}|^{\beta}, & \text{if } p_{c,x,y} = 1 \\ (1 - p_{c,x,y})^{\beta} \, \hat{p}_{c,x,y}^{\alpha} \, |p_{c,x,y} - \hat{p}_{c,x,y}|, & \text{otherwise} \end{cases} \tag{28}$$

where $p_{c,x,y}$ denotes the Gaussian-shaped soft label generated for class c at location (x, y), $\hat{p}_{c,x,y}$ is the predicted heatmap response, N is the number of objects in the image, and α, β are focusing parameters (typically $\alpha = 2$, $\beta = 4$) that suppress gradients from easy samples while emphasizing hard examples.

The size regression loss utilizes Smooth L_1 loss to measure the deviation between predicted and ground-truth bounding box dimensions:

$$\mathcal{L}_{wh} = \frac{1}{N} \sum_{k=1}^{N} \text{Smooth}_{L_1} \left(\hat{s}_k - s_k \right) \tag{29}$$

where $s_k = (w_k, h_k)$ represents the true width and height of the k-th object. This formulation provides robustness against outliers compared to L_2 loss, preventing excessive gradient updates from erroneous predictions.

The offset loss adopts L_1 loss to correct quantization errors induced by feature map downsampling:

$$\mathcal{L}_{off} = \frac{1}{N} \sum_{k=1}^{N} \left\| \hat{o}_k - o_k \right\|_1 \tag{30}$$

where $o_k = \left(\frac{x_k}{R} - \lfloor \frac{x_k}{R} \rfloor, \; \frac{y_k}{R} - \lfloor \frac{y_k}{R} \rfloor \right)$ denotes the sub-pixel offset of the k-th object center, R is the total downsampling stride, and $\hat{o}_k$ is the network prediction.

The overall training objective is formulated as a weighted sum:

$$\mathcal{L} = \lambda_{hm}\mathcal{L}_{hm} + \lambda_{wh}\mathcal{L}_{wh} + \lambda_{off}\mathcal{L}_{off} \tag{31}$$

with empirically tuned weights $\lambda_{hm} = 1$, $\lambda_{wh} = 0.1$, and $\lambda_{off} = 1$. This weighting scheme ensures that heatmap learning dominates convergence while providing stable gradients to the size and offset branches, thereby achieving pixel-accurate center localization and precise geometric estimation for robust moving object detection.

The model was trained for 70 epochs on a single NVIDIA RTX 4090 GPU with a batch size of 2. We adopted an initial learning rate of 1.25×10^{-4} with the Adam Optimizer and applied a step-wise learning rate decay strategy, reducing the learning rate by a factor of 10 at epoch 45 and 55.

4.4. Comparative Experimental Results

To further validate the performance of CASTFA, we compared the proposed method with several state-of-the-art moving object detection methods. The comparison results between CASTFA and these methods are presented in Table 1.

Table 1. Comparisons with advanced moving small-object detection methods on RSCarData.

Methods	Video 1 (ID:3)			Video 2 (ID:5)			Video 3 (ID:2)			Video 4 (ID:8)		
	Re	**Pr**	**F1**	**Re**	**Pr**	**F1**	**Re**	**Pr**	**F1**	**Re**	**Pr**	**F1**
Vibe [37]	0.61	0.34	0.44	0.82	0.61	0.70	0.68	0.59	0.63	0.65	0.52	0.58
GoDec [38]	0.92 *	0.51	0.65	0.73	0.81	0.77	0.93 *	0.53	0.68	0.72	0.38	0.50
Decolor [39]	0.24	0.92	0.38	0.77	0.88 *	0.82	0.89	0.83	0.86 *	0.44	0.93 *	0.60
ClusterNet [27]	0.75	0.67	0.71	0.66	0.81	0.72	0.90	0.72	0.80	0.50	0.70	0.58
DTTP [40]	0.74	0.67	0.70	0.67	0.84	0.74	0.71	0.84	0.77	0.64	0.86	0.73
AGMM [41]	0.72	0.56	0.63	0.80	0.77	0.79	0.93 *	0.65	0.76	0.87 *	0.62	0.72
E-LSD [1]	0.71	0.83	0.77	0.75	0.88 *	0.81	0.64	0.67	0.65	0.61	0.86	0.72
D&T [42]	0.71	0.91	0.80	0.69	0.86	0.76	0.84	0.84	0.84	0.75	0.85	0.80
B-MCMD [3]	0.77	0.93	0.85	0.76	0.86	0.81	0.86	0.82	0.84	0.71	0.77	0.74
DSFNet [19]	0.90	0.95 *	0.92 *	0.85 *	0.81	0.83 *	0.90	0.75	0.82	0.83	0.87	0.85 *
MICPL [21]	0.89	0.92	0.90	0.78	0.87	0.82	0.84	0.86 *	0.85	0.74	0.90	0.81
CASTFA	0.92 *	0.90	0.91	0.83	0.83	0.83 *	0.90	0.78	0.84	0.83	0.85	0.84

Methods	Video 5 (ID:10)			Video 6 (ID:6)			Video 7 (ID:9)			Average		
	Re	**Pr**	**F1**	**Re**	**Pr**	**F1**	**Re**	**Pr**	**F1**	**Re**	**Pr**	**F1**
Vibe [37]	0.72	0.65	0.69	0.60	0.42	0.49	0.45	0.44	0.44	0.65	0.51	0.57
GoDec [38]	0.72	0.74	0.73	0.81 *	0.42	0.55	0.93 *	0.25	0.39	0.85	0.52	0.61
Decolor [39]	0.74	0.84	0.79	0.71	0.80	0.75	0.30	0.69	0.42	0.58	0.84	0.66
ClusterNet [27]	0.76	0.82	0.79	0.77	0.71	0.74	0.85	0.66	0.75	0.74	0.73	0.73
DTTP [40]	0.62	0.77	0.69	0.55	0.73	0.62	0.25	0.49	0.33	0.60	0.74	0.65
AGMM [41]	0.76	0.68	0.72	0.79	0.53	0.63	0.90	0.37	0.53	0.82	0.60	0.68
E-LSD [1]	0.57	0.92 *	0.70	0.55	0.82	0.66	0.58	0.61	0.60	0.63	0.80	0.70
D&T [42]	0.63	0.82	0.71	0.64	0.76	0.70	0.83	0.43	0.56	0.73	0.78	0.74
B-MCMD [3]	0.58	0.84	0.68	0.70	0.74	0.72	0.81	0.47	0.60	0.77	0.77	0.76
DSFNet [19]	0.89 *	0.73	0.80	0.78	0.85	0.81 *	0.70	0.79 *	0.74	0.86 *	0.81	0.83
MICPL [21]	0.81	0.88	0.84 *	0.73	0.88 *	0.80	0.62	0.64	0.63	0.80	0.87 *	0.83
CASTFA	0.88	0.79	0.83	0.75	0.85	0.80	0.77	0.76	0.77 *	0.84	0.83	0.84 *

* Best result in each column. Underlined indicates the second-best result.

Experimental results demonstrate that the CASTFA method exhibits comprehensive and robust performance advantages in moving small-object detection on satellite videos. Systematic evaluation on the RSCarData dataset shows that CASTFA achieves a superior average F1-score of 0.84, outperforming all competing methods by a significant margin—8 percentage points higher than the second-best method (B-MCMD), and 27 and 23 percentage points above traditional approaches such as ViBe and GoDec, respectively. Notably,

CASTFA consistently attains the highest F1-score across all seven test videos while maintaining an excellent balance between recall and precision, effectively reducing both missed detections and false alarms. This consistently strong performance across diverse scenarios underscores CASTFA's statistical significance, environmental adaptability, and detection robustness, establishing it as a reliable solution for small-object detection in complex satellite video sequences.

4.5. Cross-Dataset Generalization Validation

To assess whether the performance gains of CASTFA are specific to the RSCarData benchmark or generalize to other satellite video datasets, we conduct a cross-dataset evaluation on the SAT-MTB car subset. All compared methods are retrained or re-evaluated on the SAT-MTB training split using identical hyperparameters as described in Section 4.2, and evaluated on the SAT-MTB test split.

Table 2 reports the cross-dataset comparison results. CASTFA consistently outperforms all competing methods on the SAT-MTB benchmark by a substantial margin, whereas traditional background subtraction methods such as ViBe and GoDec suffer severe performance degradation under the more challenging imaging conditions of SAT-MTB. Compared to RSCarData, the SAT-MTB test set presents considerably more stringent imaging conditions—including lower object contrast, greater background diversity spanning multiple continents and climate zones, and a higher proportion of extremely small vehicles—making it a more demanding testbed for moving vehicle detection. Despite these increased difficulties, CASTFA maintains significantly higher detection accuracy than all baselines, confirming that the multi-domain feature alignment strategy generalizes beyond the training distribution of RSCarData and is not overfitted to a specific satellite platform or scene type.

Table 2. Cross-dataset comparison results on the SAT-MTB car test set.

Method	Precision	Recall	F1
ViBe [37]	0.10	0.10	0.10
GoDec [38]	0.23	0.42	0.30
DECOLOR [39]	0.03	0.25	0.06
ClusterNet [27]	0.07	0.09	0.07
AGMM [41]	0.30	0.33	0.31
DTTP [40]	0.57	0.27	0.37
E-LSD [1]	0.46	<u>0.51</u>	0.48
B-MCMD [3]	<u>0.81</u>	0.06	0.10
DSFNet [19]	0.71	0.43	<u>0.54</u>
MICPL [21]	0.53	0.29	0.37
CASTFA	0.83 *	0.58 *	0.69 *

* Best result in each column. <u>Underlined</u> indicates the second-best result.

Despite the SAT-MTB car subset presenting several challenges absent from RSCarData, CASTFA maintains superior detection accuracy, demonstrating that the channel-attentive fusion of spatial, frequency, and temporal features provides transferable and generalizable representations for moving vehicle detection in satellite videos.

4.6. Computational Efficiency Analysis

We compare the computational efficiency of CASTFA against two representative methods that employ 3D convolutions for temporal modeling, DSFNet [19] and MICPL [21]. Table 3 reports the parameter count and inference time at two representative input resolutions on a single NVIDIA RTX 4090 GPU with a batch size of 1 and 5 temporal input frames. DSFNet achieves the highest inference speed with the fewest parameters, primarily because its lightweight 3D branch performs single-pass spatio-temporal convolution with-

out recurrent iterations. In contrast, the SSGTC module in CASTFA relies on a VMRNN whose hidden state must be updated iteratively across the temporal sequence to capture long-range motion dependencies, which inherently increases per-frame computation. Furthermore, the cross-domain feature alignment stage introduces additional channel attention modules (SFCA and STCA) that contribute non-trivial parameter overhead. Nevertheless, a comparison with MICPL—which also incorporates recurrent temporal modeling—is instructive: CASTFA contains approximately twice the parameters of MICPL, yet achieves comparable inference speed at both tested resolutions. This indicates that the selective state space mechanism in SSGTC, despite its iterative hidden state updates, avoids the quadratic complexity of conventional attention and thus remains computationally tractable. Moreover, the multi-branch aggregation design of CASTFA, though more parameter-intensive, yields substantial gains in detection accuracy and cross-dataset generalization over both DSFNet and MICPL, as demonstrated in Sections 1 and 4.5. Future efforts on model compression and cross-branch parameter sharing may further reduce the parameter footprint without compromising the benefits of multi-domain feature alignment.

Table 3. Computational efficiency comparison of deep learning methods (batch size = 1, 5 temporal frames).

Method	Params (M)	512×512	1024×1024
DSFNet [19]	19.84	36.20 ms/27.60 FPS *	135.10 ms/7.40 FPS *
MICPL [21]	18.33 *	72.40 ms/13.80 FPS	287.70 ms/3.50 FPS
CASTFA	39.57	71.80 ms/13.90 FPS	264.90 ms/3.80 FPS

* Best result in each column (fewer parameters or higher FPS). Underlined indicates the second-best result.

4.7. Ablation Study

To precisely quantify the contribution of each core component in the proposed motion object detection method, we designed ablation studies targeting key modules including multi-scale frequency representation, temporal information extraction, spatial-frequency channel attention, and spatial–temporal channel attention. By systematically removing individual modules and observing variations in detection precision (Pr), recall (Re), and F1 score, we aim to validate the necessity of each component and avoid attributing performance gains merely to increased model complexity. Table 4 summarizes the ablation results under different configurations on the RSCarData dataset, clearly illustrating the specific impact of each module on detection performance and providing empirical evidence for effectiveness analysis and structural optimization.

Table 4. Ablation study results of the proposed modules on the RSCarData dataset.

Configuration	Pr	Re	F1
w/o MSFR	0.70	0.77	0.73
w/o SSGTC	0.76	0.71	0.73
w/o SFCA	0.78	0.84 *	0.80
w/o STCA	0.76	0.79	0.78
Full model	0.84 *	0.83	0.84 *

* Best result in each column. Underlined indicates the second-best result.

The full model achieves optimal performance on the RSCarData dataset, attaining the highest F1 score among all configurations, which demonstrates that the synergistic operation of the four core components effectively enhances both the accuracy and completeness of motion object detection. Removing the multi-scale frequency representation leads to the most significant drop in recall, indicating that this module substantially improves the perception of multi-scale motion targets by fusing motion responses across different frequency

bands, thereby effectively reducing missed detections, especially for weak motions or small objects. The absence of the temporal information extraction module causes synchronous declines in precision and F1 score, reflecting its capability to model inter-frame motion consistency and suppress false positives triggered by static background disturbances or transient noise. When the spatial-frequency channel attention is removed, precision noticeably deteriorates, suggesting that this mechanism adaptively weights frequency-domain feature channels to reinforce spectral components with strong motion discriminability. The removal of spatial–temporal channel attention primarily affects recall, revealing its critical role in associating spatial features across frames to maintain temporal continuity of detected objects and prevent target loss caused by transient occlusion or motion blur. Figure 8 visually compares the detection results under different configurations. As shown in the figure, the full model produces detection outputs closest to the ground truth, whereas other configurations exhibit varying degrees of missed detections and false alarms—for instance, misidentifying objects with vehicle-like shapes or colors as targets, or overlooking vehicles with low contrast against the background. In summary, the four modules collectively enhance detection performance from four complementary perspectives: frequency-domain representation, temporal modeling, frequency-aware feature selection, and spatiotemporal feature integration.

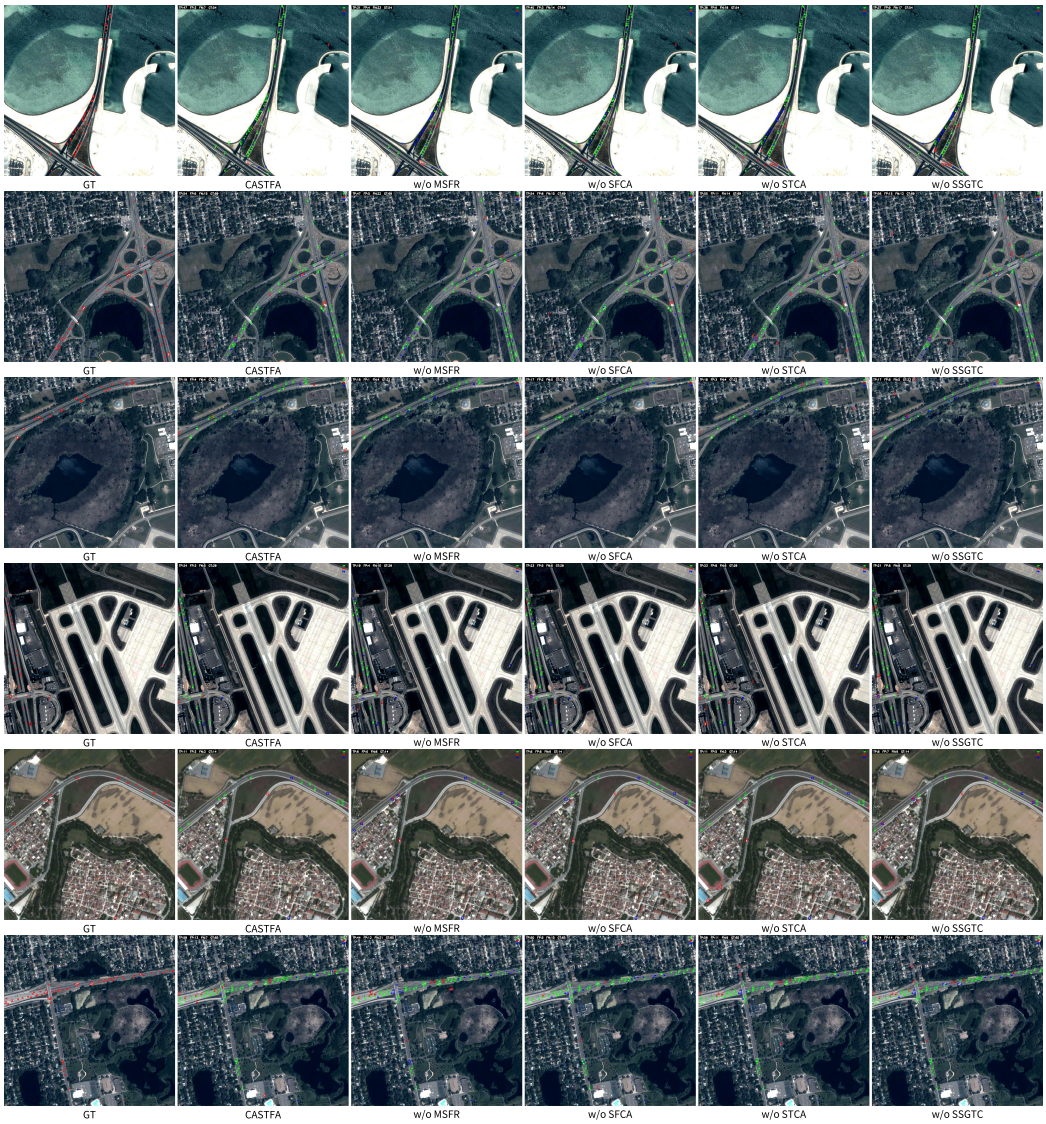


Figure 8. Visual comparison of ablation study results.

To demonstrate that CASTFA effectively leverages temporal information from video sequences, we design a progressive temporal ablation that incrementally incorporates past frames into the model input. Specifically, five consecutive frames are sampled from sequence 02 of RSCarData, and past frames are progressively added to the temporal window ($T = 1$ to $T = 5$), with the remaining positions padded by the current frame t. Figure 9 visualizes the results under five temporal configurations (columns). The first row shows the input frame stack for each configuration. The second row displays the frame difference between each past frame and the current frame t, revealing the inter-frame motion cues available to the network. The third row presents the output heatmaps predicted by CASTFA under each setting, and the fourth row highlights the heatmap variation induced by incorporating each additional past frame. As more past frames are introduced, the heatmap responses on moving vehicles become progressively stronger and more concentrated, while static background regions remain largely unaffected. This confirms that the temporal branch (SSGTC) effectively extracts inter-frame motion cues and translates them into enhanced detection confidence for moving objects. The progressive nature of the improvement also corroborates the quantitative ablation results in Table 4, where removing SSGTC leads to the largest drop in detection performance.

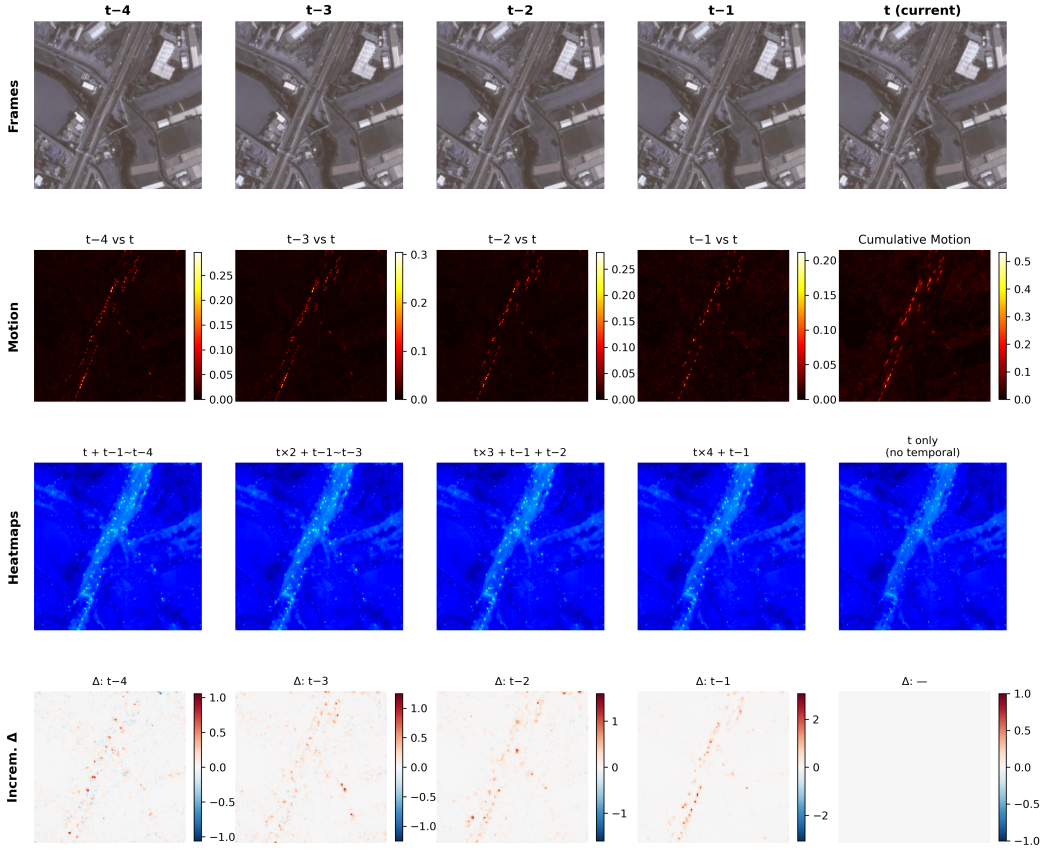


Figure 9. Progressive temporal information visualization on sequence 02 of RSCarData. From left to right, $T = 1$ to $T = 5$ past frames are progressively incorporated into the input (remaining positions padded with the current frame t). From top to bottom: input frame stack, frame difference w.r.t. the current frame, predicted heatmap, and heatmap variation induced by the newly added past frame. The heatmap responses intensify as more temporal context becomes available, demonstrating the network's ability to exploit inter-frame motion cues for detection.

5. Discussion

Compared with existing methods for moving object detection in remote sensing videos, the core advantage of CASTFA lies in its ability to collaboratively model multi-

domain features. Conventional approaches often rely on feature extraction within a single domain—such as using spatial convolutions solely for appearance modeling or simple temporal stacking for motion cues—making them ill-suited to handle challenges inherent in remote sensing scenarios, including multi-scale targets, complex backgrounds, and irregular motion patterns. In contrast, CASTFA achieves multi-perspective representation of moving objects through parallel perception across spatial, frequency, and temporal domains: the spatial domain captures geometric structures and local details of vehicles; the frequency domain decomposes image content into low-frequency approximations and high-frequency directional components, which helps distinguish motion-induced spectral changes from static background textures; and the temporal domain leverages a selective state space model to capture long-range motion dependencies with linear computational complexity, enabling more stable trajectory modeling than conventional recurrent networks.

The multi-domain collaborative design of CASTFA offers potential value for urban traffic monitoring applications. Traffic flow estimation in smart city planning requires continuous, large-scale vehicle detection data with reasonable accuracy and temporal consistency. The integration of frequency-domain features may contribute to more robust detection under challenging imaging conditions commonly encountered in urban remote sensing, such as pronounced texture variation across different land-use types and dynamic scene complexity at major road intersections. The temporal modeling capacity of SSGTC supports the analysis of motion patterns over extended video sequences, which is relevant for studying traffic dynamics including congestion formation, vehicle queuing behavior, and intersection throughput under varying demand conditions. While these capabilities suggest that CASTFA could serve as a component in urban traffic data acquisition pipelines, its practical utility for planning and governance decisions would depend on further validation across diverse city environments and integration with downstream analytical frameworks.

Furthermore, the channel attention-driven multi-domain fusion mechanism adaptively regulates the contribution of each feature domain based on input content, avoiding the rigidity of fixed fusion strategies. This adaptability allows the model to adjust its reliance on spatial, frequency, and temporal cues according to scene conditions. For example, when motion cues are weak or ambiguous, the model can increase the weight of frequency-domain and spatial-domain features to maintain detection reliability; conversely, when vehicles exhibit clear temporal motion patterns, the model can emphasize temporal-domain information to improve tracking consistency. Such content-aware feature selection provides a more flexible foundation for handling the heterogeneous visual conditions present in satellite video coverage of urban areas, where scene characteristics may vary substantially across different city districts and time periods.

Nevertheless, several architectural limitations of CASTFA should be acknowledged. First, the parallel three-branch structure introduces significant computational redundancy, as each branch performs independent feature extraction and maintains separate parameter sets. This design choice increases both inference time and model size, which may constrain deployment scenarios requiring real-time or high-throughput processing of extensive video data. Second, the dynamic cross-domain fusion operations involve feature transformation and normalization steps that elevate memory consumption during inference. For applications involving simultaneous monitoring of multiple spatial regions or extended video sequences, this memory overhead could limit batch processing capacity or necessitate higher-end hardware configurations. Third, despite the theoretical linear complexity of the selective state space model, its implementation relies on sequential scanning operations and input-dependent parameter generation that may not fully exploit the parallel computa-

tion capabilities of modern GPU architectures, potentially affecting training efficiency and inference speed when scaling to larger datasets.

Future research directions include investigating model compression and architecture optimization techniques to reduce the computational and memory footprint of the multi-branch framework. Possible approaches include branch parameter sharing, knowledge distillation, or neural architecture search to identify more efficient configurations. Additionally, integrating CASTFA with downstream urban analytics workflows—such as traffic flow aggregation, congestion prediction, and origin-destination estimation—would help establish the practical value of detection outputs for transportation planning applications. Exploration of lightweight deployment strategies, including edge computing and quantized inference, may also facilitate the transition from research prototypes to operational components of city-scale intelligent transportation systems.

6. Conclusions

This study proposes CASTFA, a novel method for moving object detection in remote sensing videos, which achieves deep integration of multi-dimensional target representations through collaborative modeling of spatial-, frequency-, and temporal-domain features. The core innovation of CASTFA lies in its parallel multi-domain perception pathways and a channel attention-driven adaptive fusion mechanism, enabling the model to dynamically allocate feature contributions from different domains according to scene-specific characteristics. This design endows the method with improved robustness and detection stability under challenging conditions such as complex background clutter, significant scale variations, and irregular motion patterns. Experimental results demonstrate that CASTFA effectively enhances the spatiotemporal consistency of moving object representations by leveraging the spatial domain to capture geometric structures, the frequency domain to provide complementary spectral information, and the temporal domain to model motion dynamics.

The accurate moving vehicle detection enabled by CASTFA contributes to the acquisition of large-scale traffic flow data from satellite video sources, which could potentially support urban traffic monitoring, congestion analysis, and transportation planning workflows. However, realizing this potential in operational smart city systems will require addressing the computational efficiency and scalability challenges identified in this work. Future research will focus on model compression and architecture optimization to reduce inference costs, as well as integration with urban analytical platforms to bridge the gap between detection outputs and actionable planning insights.

Author Contributions: Conceptualization, Ning Zhao, Jun Shi and Haopeng Zhang; methodology, Ning Zhao, Xiao Wang and Xiaopeng Zhang; software, Ning Zhao, Xiao Wang and Xiaopeng Zhang; validation, Ning Zhao, Xiao Wang, Xiaopeng Zhang and Jun Shi; formal analysis, Ning Zhao, Xiao Wang and Xiaopeng Zhang; investigation, Ning Zhao, Xiao Wang, Xiaopeng Zhang and Jun Shi; resources, Zhiguo Jiang and Haopeng Zhang; data curation, Xiao Wang and Xiaopeng Zhang; writing—original draft preparation, Ning Zhao, Xiao Wang and Xiaopeng Zhang; writing—review and editing, Jun Shi, Zhiguo Jiang and Haopeng Zhang; visualization, Ning Zhao, Xiao Wang and Xiaopeng Zhang; supervision, Zhiguo Jiang and Haopeng Zhang; project administration, Jun Shi, Zhiguo Jiang and Haopeng Zhang; funding acquisition, Jun Shi, Zhiguo Jiang and Haopeng Zhang. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by the Major Science and Technology Program of Zhejiang Province (Grant No. 2026LDC01007(JT)), the Key Research Program of Hangzhou (Grant No. 2025SZD2B02), and the Open Research Fund of Anhui Key Laboratory of Optical Quantitative Remote Sensing (No. OQRS-KFJJ-01).

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

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

References

1. Zhang, J.; Jia, X.; Hu, J.; Huang, Q. Error-Bounded Foreground and Background Modeling for Moving Object Detection in Satellite Videos. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 2659–2669. [CrossRef]
2. Zhang, J.; Jia, X.; Hu, J.; Tan, S.; Huang, Q. Online Structured Sparsity-Based Moving-Object Detection from Satellite Videos. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 6420–6433. [CrossRef]
3. Zhang, J.; Jia, X.; Hu, J.; Tan, K. Moving Vehicle Detection for Remote Sensing Video Surveillance with Nonstationary Satellite Platform. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 5185–5198. [CrossRef] [PubMed]
4. Zhang, X.; Xiang, J.; Zhang, Y. Space Object Detection in Video Satellite Images Using Motion Information. *Int. J. Aerosp. Eng.* **2017**, *2017*, 1024529. [CrossRef]
5. Li, H.; Chen, L.; Li, F.; Huang, M. Ship detection and tracking method for satellite video based on multiscale saliency and surrounding contrast analysis. *J. Appl. Remote Sens.* **2019**, *13*, 026511. [CrossRef]
6. Shi, F.; Qiu, F.; Li, X.; Zhong, R.; Yang, C.; Tang, Y. Detecting and tracking moving airplanes from space based on normalized frame difference labeling and improved similarity measures. *Remote Sens.* **2020**, *12*, 3589. [CrossRef]
7. Chen, X.; Sui, H.; Fang, J.; Zhou, M.; Wu, C. A Novel AMS-DAT Algorithm for Moving Vehicle Detection in a Satellite Video. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
8. Cao, D.; Ma, J.; Chen, Z. Video object detection algorithm based on dynamic combination of sparse feature propagation and dense feature aggregation. *Multimed. Tools Appl.* **2021**, *80*, 23275–23295. [CrossRef]
9. Li, H.; Chen, L.; Li, F. Moving Target Detection Based on Edge Optical Flow from Satellite Videos. In Proceedings of the Optoelectronic Imaging and Multimedia Technology V, Beijing, China, 11–12 October 2018; Volume 10817, pp. 114–119. [CrossRef]
10. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
11. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
12. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
13. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic DETR: End-to-End Object Detection with Dynamic Attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 2968–2977. [CrossRef]
14. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6568–6577. [CrossRef]
15. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 765–781. [CrossRef]
16. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet++ for Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 3509–3521. [CrossRef] [PubMed]
17. Chen, R.; Li, X.; Li, S. A Lightweight CNN Model for Refining Moving Vehicle Detection From Satellite Videos. *IEEE Access* **2020**, *8*, 221897–221917. [CrossRef]
18. Yao, C.H.; Fang, C.Y.; Shen, H.H.; Hsiao, P.Y. Video Object Detection via Object-Level Temporal Aggregation. In *European Conference on Computer Vision*; Springer International Publishing: Cham, Switzerland, 2020; pp. 160–177. [CrossRef]
19. Xiao, C.; Yin, Q.; Ying, X.; Li, R.; Wu, S.; Li, M.; Liu, L.; An, W.; Chen, Z. DSFNet: Dynamic and Static Fusion Network for Moving Object Detection in Satellite Videos. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 3510405. [CrossRef]
20. Zhu, X.; Li, J.; Feng, J.; Xia, L.; Shen, H. 3D-Mglnet: Moving Vehicle Detection in Satellite Videos with 3D Motion-Guided Lightweight Network. In Proceedings of the IGARSS 2023–2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; pp. 6478–6481. [CrossRef]
21. Chen, S.; Ji, L.; Zhu, S.; Ye, M. MICPL: Motion-Inspired Cross-Pattern Learning for Small-Object Detection in Satellite Videos. *IEEE Trans. Neural Netw. Learn. Syst.* **2025**, *36*, 6437–6450. [CrossRef] [PubMed]
22. Feng, J.; Jiang, Q.; Zhang, J.; Liang, Y.; Shang, R.; Jiao, L. CFDRM: Coarse-to-Fine Dynamic Refinement Model for Weakly Supervised Moving Vehicle Detection in Satellite Videos. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5626413. [CrossRef]

23. Kopsiaftis, G.; Karantzalos, K. Vehicle detection and traffic density monitoring from very high resolution satellite video data. In Proceedings of the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 26–31 July 2015; pp. 1881–1884. [CrossRef]
24. Zhang, X.; Feng, Y.; Wang, N.; Lu, G.; Mei, S. Aerial Person Detection for Search and Rescue: Survey and Benchmarks. *J. Remote Sens.* **2025**, *5*, 0474. [CrossRef]
25. Zhang, X.; Feng, Y.; Zhang, S.; Wang, N.; Lu, G.; Mei, S. Robust Aerial Person Detection with Lightweight Distillation Network for Edge Deployment. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5630616. [CrossRef]
26. Zhang, X.; Feng, Y.; Zhao, Z.; Ren, Y.; Mei, S. Rethinking Coarse-to-Fine Fully Convolutional Transformers for Salient Maritime Ship Detection in Optical Remote Sensing Imagery. *J. Remote Sens.* **2026**, *6*, 1038. [CrossRef]
27. LaLonde, R.; Zhang, D.; Shah, M. ClusterNet: Detecting Small Objects in Large Scenes by Exploiting Spatio-Temporal Information. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4003–4012. [CrossRef]
28. Feng, J.; Zeng, D.; Jia, X.; Zhang, X.; Li, J.; Liang, Y.; Jiao, L. Cross-frame keypoint-based and spatial motion information-guided networks for moving vehicle detection and tracking in satellite videos. *ISPRS J. Photogramm. Remote Sens.* **2021**, *177*, 116–130. [CrossRef]
29. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
30. Yu, F.; Wang, D.; Shelhamer, E.; Darrell, T. Deep Layer Aggregation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2403–2412. [CrossRef]
31. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
32. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2023**, arXiv:2312.00752. [CrossRef]
33. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 802–810. [CrossRef]
34. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
35. Yin, Q.; Hu, Q.; Liu, H.; Zhang, F.; Wang, Y.; Lin, Z.; An, W.; Guo, Y. Detecting and Tracking Small and Dense Moving Objects in Satellite Videos: A Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5612518. [CrossRef]
36. Li, S.; Zhou, Z.; Zhao, M.; Yang, J.; Guo, W.; Lv, Y.; Kou, L.; Wang, H.; Gu, Y. A Multitask Benchmark Dataset for Satellite Video: Object Detection, Tracking, and Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5611021. [CrossRef]
37. Barnich, O.; Van Droogenbroeck, M. ViBe: A Universal Background Subtraction Algorithm for Video Sequences. *IEEE Trans. Image Process.* **2011**, *20*, 1709–1724. [CrossRef] [PubMed]
38. Zhou, T.; Tao, D. GoDec: Randomized low-rank & sparse matrix decomposition in noisy case. In Proceedings of the 28th International Conference on International Conference on Machine Learning, Madison, WI, USA, 28 June–2 July 2011; ICML11; pp. 33–40. [CrossRef]
39. Zhou, X.; Yang, C.; Yu, W. Moving Object Detection by Detecting Contiguous Outliers in the Low-Rank Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 597–610. [CrossRef] [PubMed]
40. Ahmadi, S.A.; Ghorbanian, A.; Mohammadzadeh, A. Moving vehicle detection, tracking and traffic parameter estimation from a satellite video: A perspective on a smarter city. *Int. J. Remote Sens.* **2019**, *40*, 8379–8394. [CrossRef]
41. Lin, H.H.; Chuang, J.H.; Liu, T.L. Regularized Background Adaptation: A Novel Learning Rate Control Scheme for Gaussian Mixture Modeling. *IEEE Trans. Image Process.* **2011**, *20*, 822–836. [CrossRef] [PubMed]
42. Ao, W.; Fu, Y.; Hou, X.; Xu, F. Needles in a Haystack: Tracking City-Scale Moving Vehicles From Continuously Moving Satellite. *IEEE Trans. Image Process.* **2020**, *29*, 1944–1957. [CrossRef] [PubMed]

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