HTTrack: Learning to Perceive Targets via Historical Trajectories in Satellite Video Tracking

Jiahao Wang[1, 2, 3], Fang Liu[1, 2, 3]*, Licheng Jiao[1, 2, 3], Hao Wang[1, 2, 3], Shuo Li[1, 2, 3], Xinyi Wang[1, 2, 3], Lingling Li[1, 2, 3], Puhua Chen[1, 2, 3], Xu Liu[1, 2, 3]

[1]Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education
[2]Joint International Research Laboratory of Intelligent Perception and Computation
[3]School of Artificial Intelligent, Xidian University, Xi'an,710071, P.R. China
jh_wang1024@163.com, f63liu@163.com

Abstract

In recent years, the rapid progress of deep learning has driven notable advancements in satellite video tracking, a critical task for applications such as environmental monitoring, disaster management, and defense. Despite these strides, existing approaches remain constrained by their inability to handle dynamic challenges, such as target appearance variations, complex motion patterns, and occlusions. Traditional methods often suffer from static template matching or overly complex update mechanisms, compromising their robustness and practicality in real-world scenarios. To address these limitations, we propose a paradigm shift in satellite video tracking by integrating historical trajectory knowledge with visual features. This fusion enhances the tracker's perceptual understanding of targets over time, enabling more adaptive and resilient tracking. By aligning spatial, temporal, and cross-modal information, our approach effectively bridges the gap between fragmented observations and coherent tracking performance, even under challenging conditions like small target detection and cluttered backgrounds. Extensive experiments conducted on multiple satellite video tracking benchmarks demonstrate the superiority of our method, with HTTrack achieving success rates of 51.5% on SV248S, 52.9% on SatSOT, and 32.6% on VISO, significantly outperforming state-of-the-art trackers, marking a step forward in achieving robust, accurate, and scalable satellite video tracking.

Introduction

Satellite video tracking, an essential branch of remote sensing technology, provides unprecedented capabilities for acquiring, monitoring, and tracking specific moving targets on Earth's surface, such as vehicles, ships, and aircraft (Zhang and Zhang 2022; Jiao et al. 2023; Pang et al. 2023; Liu et al. 2025). The advent of high-resolution satellite video systems has transformed traditional static remote sensing into a dynamic and continuous monitoring process, offering valuable insights for various applications, including military surveillance, traffic management, and environmental monitoring. As technology advances and the range of applications expands, satellite video tracking technology is poised to play an increasingly significant role in the future.

*Corresponding author

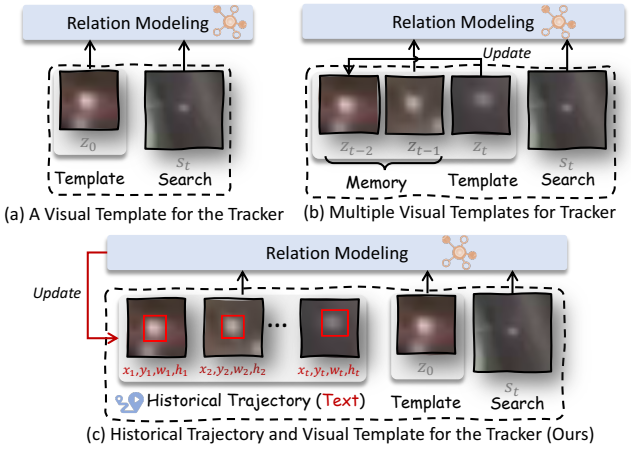

Figure 1: Visual template-based trackers and the proposed HTTrack. (a) Tracker based on a single visual template. (b) Tracker based on multiple updatable visual templates. (c) The proposed HTTrack tracker.

In recent years, the rapid development of deep learning technology has led to significant advances in object tracking, establishing deep trackers as the dominant technology in this field. Two main types of deep tracking methods are widely studied. One type treats object tracking as a similarity matching problem between a target template and search frames, constructing an embedding space during offline training. Siamese trackers represent this template matching approach (He et al. 2018; Liang and Shen 2019; Chen et al. 2020; Jiao et al. 2021b; Li et al. 2023b), as illustrated in Figure 1(a). However, these methods typically do not update the template, making it challenging to adapt to changes in the target's appearance due to occlusion or non-rigid deformation. Some trackers incorporate complex template update mechanisms to address this limitation (Yang et al. 2023; Zhang et al. 2023; Shao et al. 2024; Zhou et al. 2024), as shown in Figure 1(b). These mechanisms enable the tracker to handle appearance changes better, resulting in greater robustness than traditional Siamese trackers. However, it is important to note that these customized update

strategies often introduce hyperparameters that require intricate tuning, complicating the model's design and debugging. Consequently, the generalizability and scalability of these methods in practical applications are somewhat constrained.

While deep trackers have made remarkable progress in handling target appearance and motion variations, they still encounter significant challenges in complex scenarios. Issues such as target loss during occlusions, rapid motion, or abrupt lighting changes persist as key bottlenecks. Traditional methods often rely on intricate template update mechanisms, introducing additional hyperparameter tuning, increasing system complexity, and reducing reliability in real-world applications. To address these challenges and push the boundaries of tracking performance, we propose a novel framework that leverages large language models (LLMs) to incorporate historical trajectory information. Unlike conventional approaches, LLMs contextualize target behavior by capturing movement trends and patterns over time. This enables more accurate and robust position prediction for the current frame (Wang et al. 2024c,b). The key advantages of our approach are as follows: (I) **Enhanced Continuity and Stability**: When visual tracking suffers from temporary target loss due to occlusions, rapid movement, or lighting changes, LLM-based trajectory prediction provides critical auxiliary information, ensuring smoother and more stable tracking. (II) **Dynamic Strategy Adaptation**: By analyzing historical trajectory patterns, the model can dynamically adjust tracking strategies to better align with the target's actual movement, improving overall flexibility and responsiveness. (III) **Fusion of Complementary Information**: Visual features and historical trajectories represent two distinct yet complementary information sources. LLMs effectively integrate these modalities, mitigating the weaknesses of single-source tracking. (IV) **Maintained Tracking Consistency**: Incorporating trajectory history reinforces consistency across frames, reducing the impact of errors from individual frame predictions and preventing cascading failures.

In this work, we represent historical trajectories as textual input and combine them with visual features to enhance the perception capability of LLMs for target tracking, as shown in Figure 1(c). The proposed HTTrack leverages trajectory-aware text prompts to help predict target positions in the current frame, improving accuracy, robustness, and interpretability. To bridge the gap between modalities, we introduce two lightweight modules: the **D**ynamic **S**caling **A**lignment (**DSA**) for adaptive feature adjustment, and the **C**ross-**M**odal **A**lignment (**CMA**) for effective fusion of text and visual features. Additionally, to address the challenge of small object tracking and scale variation, we propose the **C**ross-**S**patial **F**eature **M**apper (**CSFM**), which aggregates multi-scale spatial features to preserve fine-grained target details. Overall, our method enables more reliable and context-aware satellite video tracking in complex scenarios. Our key contributions are summarized as follows:

1. We propose the **HTTrack** tracker, which integrates LLMs and visual features. Using historical trajectory texts significantly enhances the robustness and adaptability of satellite video tracking.
2. We design the **D**ynamic **S**cale **A**lignment (**DSA**) and **C**ross-**M**odal **A**lignment (**CMA**) modules, which dynamically adjust feature alignment and seamlessly integrate cross-modal information. These modules enhance the tracker's robustness and task adaptability.
3. To tackle the challenge of tracking small targets in satellite videos, we introduce the **C**ross-**S**patial **F**eature **M**apper (**CSFM**) module. This module effectively resolves issues related to small target scales, enhancing the tracker's stability and precision in complex scenarios.
4. Extensive experiments on the **SV248S**, **SatSOT**, and **VISO** datasets demonstrate that our tracker consistently outperforms existing methods in various scenarios.

Related Work

Large Language Models: In recent years, Large Language Models (LLMs) have garnered significant attention for their extensive applications in the field of Natural Language Processing (NLP). Notably, LLMs such as LLaMA (Touvron et al. 2023), MOSS (Sun et al. 2024), GLM (Du et al. 2021), OpenAI's GPT series (Brown et al. 2020), Google's BERT (Devlin et al. 2018), and T5 (Chung et al. 2024) have successfully captured and modelled long-range dependencies in text by introducing advanced self-attention mechanisms. This has significantly enhanced the models' ability to understand and generate complex semantics. These models have demonstrated exceptional performance across various NLP tasks, including text generation, machine translation, and question-answering systems, thereby pushing the boundaries of NLP technology and offering new methodologies and insights for research in other fields (Wu et al. 2023b; He et al. 2024; Yin et al. 2024; Wang et al. 2024c; Yang et al. 2025). For example, (Wang et al. 2024d) proposed an innovative multimodal LLM framework designed to leverage the powerful capabilities of LLM to address challenges in traditional vision tasks. These tasks include object detection, image segmentation, image description generation, etc., providing new perspectives on the deep integration of vision and language. (Li et al. 2024a) further advances this trend by constructing the VisionGraph benchmark dataset to explore the potential of advanced large-scale multimodal models to solve complex multimodal graph theory problems. It also introduces the concept of the description-program-reasoning (DPR) chain to enhance logical rigor in the reasoning process. Additionally, a series of models that integrate visual and language processing capabilities, such as Visual ChatGPT (Wu et al. 2023a), LLaVa (Liu et al. 2023), HuggingGPT (Shen et al. 2024), miniGPT-4 (Zhu et al. 2023), mPLUG-owl (Ye et al. 2023), and VideoChat (Li et al. 2023a), have demonstrated the feasibility of combining LLMs with visual models to solve AI tasks collaboratively. These works not only expand the application scenarios of LLMs but also lay a solid foundation for the realization of more intelligent and comprehensive AI systems.

Satellite Video Tracking: Satellite video tracking, as a crucial remote sensing technology, aims to acquire, monitor, and track specific moving targets on the Earth's surface, such as vehicles, ships, and aircraft, through satellite systems.

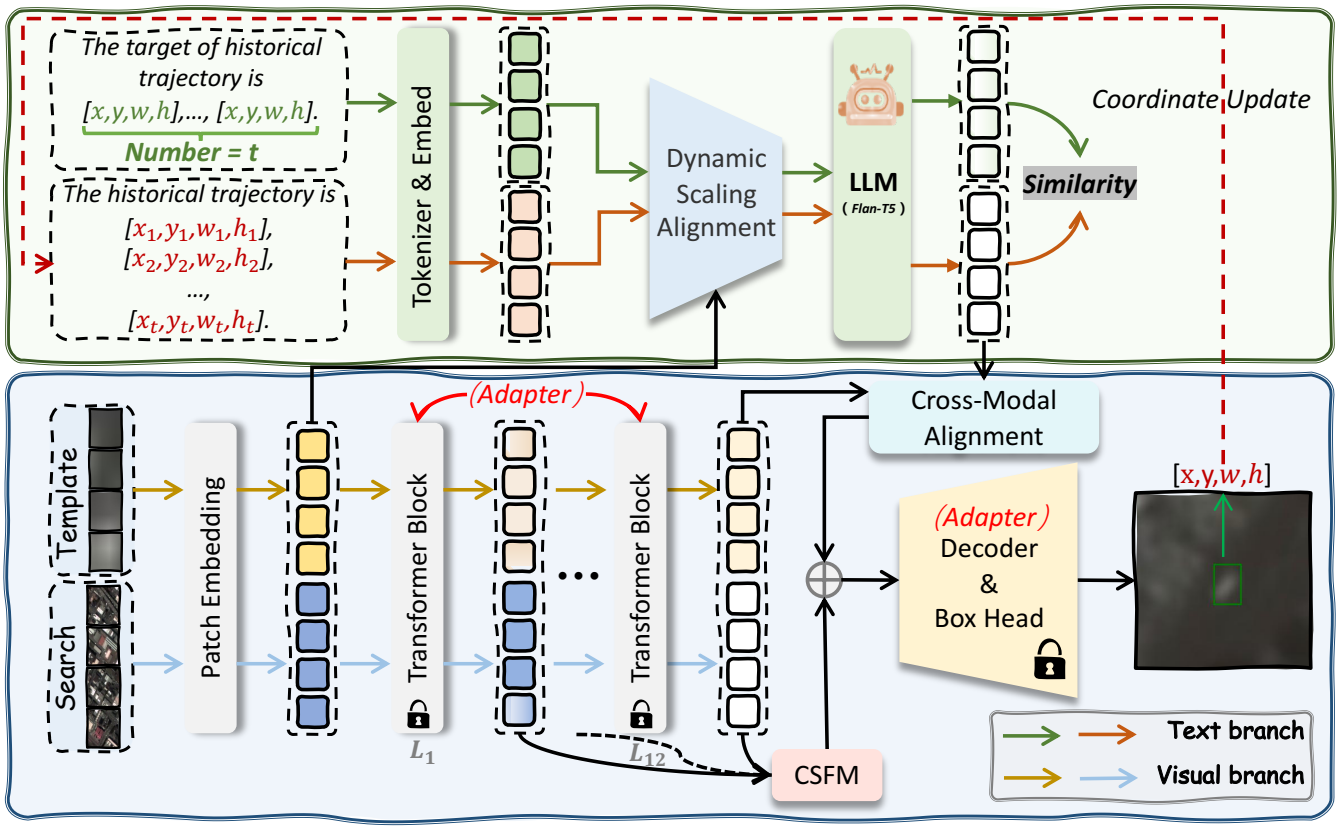

Figure 2: Overall structure of the proposed HTTrack. It consists of a text branch and a visual branch. The **red** dotted lines indicate that the historical trajectory text is continuously updated according to the predicted coordinates of the current frame during the tracking process.

During the past decade, methods based on correlation filters (Xuan et al. 2019; Zhang et al. 2020a; Jiao et al. 2021a; Chen et al. 2024; Liu et al. 2025) have dominated the field of satellite video tracking. Among these, MOSSE (Bolme et al. 2010) pioneered the foundations for subsequent research, while KCF (Henriques et al. 2014) advanced the technology by introducing kernel methods to handle non-linear features effectively. Despite the aforementioned methods alleviating some difficulties in satellite video tracking, their performance remains limited when faced with extreme environments such as occlusion, severe illumination changes, and background jitter. In recent years, significant breakthroughs have been made in the field of satellite video tracking (Zhang et al. 2021; Jiao et al. 2023; Chen et al. 2024; Huang et al. 2024; Zhong, Fang, and Shu 2024), due to the rapid development of computer vision and deep learning technologies. (Shao et al. 2021) incorporates historical image frames into the SiamRPN (Li et al. 2018) framework, utilizing Gaussian Mixture Models (GMM) (Zhang et al. 2017) and Meanshift (Shang et al. 2019) algorithms to achieve pixel-level motion target detection and tracking, significantly enhancing the comprehensive capabilities of tracking and detection. (Yang et al. 2023) focuses on addressing occlusion and background noise issues by optimizing network structures and implementing dynamic visual template update strategies, proposing an efficient SiamMDM tracking method. (Lai et al. 2023) proposes the Target-Aware Transformer method to tackle the challenges of weak small-target features and strong background interference. By introducing a Bi-Direction Propagation and Fusion strategy along with a Target-Aware Enhancement (TAE) module, their approach effectively improves tracking performance under complex scenes. In addition, (Wang et al. 2024a) builds upon the STARK (Yan et al. 2021) framework, introducing frame differencing to extract motion information, supplemented by a location hint mechanism and dynamic template updates, significantly boosting the adaptability of the tracker.

However, the dynamic visual template update strategies in existing methods often rely on manually tuned sensitive parameters, increasing operational complexity and potentially leading to significant time costs. To address this challenge, this work proposes an innovative satellite video tracking method, HTTrack, which deeply integrates the advantages of historical trajectory information and visual templates. Specifically, through the deep understanding and utilization of the target's historical trajectory by large language models (LLMs), HTTrack can provide critical auxiliary information when visual features are blurry or missing, ensuring the continuity and stability of tracking. Additionally, by combining real-time analysis of historical trajectories, HT-

Track can dynamically adjust tracking strategies to adapt to the dynamic changes in target motion patterns, significantly enhancing the adaptability and flexibility of the tracking.

Methodology

Overall Architecture

As shown in Figure 2, the proposed tracker, HTTrack, primarily consists of two branches: a text branch for extracting latent information from historical trajectories and a visual branch for extracting visual features from template and search images. This structure differs significantly from other tracking methods, primarily relying on visual templates. Given a satellite video with multiple frames, a tracker based on a visual template can be represented as $\boldsymbol{F}_V : \boldsymbol{Z}_0, \boldsymbol{S}_t \rightarrow \boldsymbol{B}$, where it predicts the target bounding box B in the current frame. For trackers that update the visual template, the model's input changes, and the representation becomes $\boldsymbol{F}_{MV} : (\boldsymbol{Z}_{t-1}, \boldsymbol{Z}_t), \boldsymbol{S}_t \rightarrow \boldsymbol{B}$. In the proposed HTTrack tracker, due to the inclusion of historical trajectories, the representation is $\boldsymbol{F}_{TV} : (\boldsymbol{T}_{t-5}, \ldots, \boldsymbol{T}_t), \boldsymbol{Z}_0, \boldsymbol{S}_t \rightarrow \boldsymbol{B}$, where $\boldsymbol{Z}_0, \boldsymbol{Z}_t \in \mathbb{R}^{3\times H_z \times W_z}$ denote the visual templates, $(\boldsymbol{T}_{t-5}, \ldots, \boldsymbol{T}_t)$ represents the continuously updated historical trajectories composed of text and $\boldsymbol{S}_t \in \mathbb{R}^{3\times H_s \times W_s}$ denotes the search images. The details of the text and visual branches are described below:

Text Branch. During the training phase, HTTrack's text branch takes the historical trajectory text $\boldsymbol{T}$ = "*The historical trajectory is* $[x_1, y_1, w_1, h_1]$, $[x_2, y_2, w_2, h_2]$, $\ldots$, $[x_t, y_t, w_t, h_t]$." and the label text of the current frame $\boldsymbol{T}_B$ = "*The target of historical trajectory is* $[x, y, w, h]$,...,$[x, y, w, h]$." as inputs. $\boldsymbol{T}$ and $\boldsymbol{T}_B$ are initially mapped to text embedding tokens $\boldsymbol{X}_T$ and $\boldsymbol{X}_{T_B}$, respectively, where $\boldsymbol{X}_T, \boldsymbol{X}_{T_B} \in \mathbb{R}^{N_T \times D}$, N_T represents the number of text tokens, and D represents the embedding dimension. These text embedding tokens and the visual feature tokens from the visual branch are input into the DSA module for dynamic scaling and fusion, resulting in cross-modal joint features. The transformed feature tokens have the same dimensions as the text embedding tokens. Finally, the associated $\boldsymbol{T}$ and $\boldsymbol{T}_B$ with visual feature tokens are fed into the LLM, generating the final language token embeddings $\boldsymbol{X}'_T$ and $\boldsymbol{X}'_{T_B}$. By constraining $\boldsymbol{X}'_T$ with $\boldsymbol{X}'_{T_B}$, features more relevant to the current frame's target is produced.

Visual Branch. Similarly to the text branch, the given template and search images are first encoded into feature representations $\boldsymbol{X}_V \in \mathbb{R}^{N\times D}$, then passed through 12 layers of Transformer Blocks for deep feature interaction, resulting in the output $\boldsymbol{X}'_V \in \mathbb{R}^{N\times D}$. Subsequently, $\boldsymbol{X}'_V$ undergoes cross-modal alignment with the historical trajectory tokens from the text branch via the CMA module. The CSFM module further enhances $\boldsymbol{X}'_V$ by incorporating outputs from different layers of the Transformer Blocks, compensating for potential missing details and enriching the final feature representation $\boldsymbol{X}_{fuse} \in \mathbb{R}^{N\times D}$. Finally, $\boldsymbol{X}_{fuse}$ is processed by the same decoder and prediction head as in (Chen et al. 2023). It is important to note that during the training phase, all parameters of the Transformer Blocks, decoder, and prediction head in the visual branch are frozen, with an Adapter added for optimization (details provided in the Appendix), allowing HTTrack to better adapt to the complex scenarios.

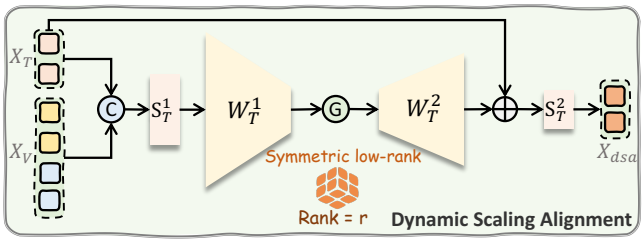

Figure 3: Detailed structure of DSA. Here, "C" represents the concatenation operation, and "G" denotes the GELU activation function.

Dynamic Scaling Alignment

To effectively align the visual features X_V with the historical trajectory X_T and ensure that the aligned token dimensions are consistent with $\boldsymbol{X}_T$, we propose **D**ynamic **S**caling **A**lignment (**DSA**), as illustrated in Figure 3. We use two learnable vectors, $\boldsymbol{S}_T^1 \in \mathbb{R}^{1\times(N+N_T)}$ and $\boldsymbol{S}_T^2 \in \mathbb{R}^{1\times N_T}$, as scaling factors to adjust the feature tokens before and after the change in the number of tokens. These learnable parameters adaptively balance token contributions, ensuring consistent feature representation and seamless integration of diverse information. This dynamic scaling enhances robustness and alignment across varying scenarios, particularly in cross-modal and temporal contexts. The specific transformation process is as follows:

$$f' = Concat(\boldsymbol{X}_T, \boldsymbol{X}_V) \cdot \boldsymbol{S}_T^1, \tag{1}$$

$$f = GELU(f' \cdot \boldsymbol{W}_T^1) \cdot \boldsymbol{W}_T^2, \tag{2}$$

$$\boldsymbol{X}_{dsa} = (\boldsymbol{X}_T + f) \cdot \boldsymbol{S}_T^2, \tag{3}$$

where $Concat(\cdot,\cdot)$ denotes concatenation along the token dimension. The matrices $\boldsymbol{W}_T^1 \in \mathbb{R}^{(N+N_T)\times r}$ and $\boldsymbol{W}_T^2 \in \mathbb{R}^{r\times N_T}$ are symmetric low-rank matrices implemented via linear layers, which handle the features before and after scaling, with $r \ll N, N_T$.

Cross-Spatial Feature Mapper

Satellite videos often cover wide areas and exhibit significant variations in target scale, where targets are typically small, making cross-scale and cross-spatial feature extraction highly effective in relevant fields. However, due to RGB-based pre-trained foundation models, these feature representations may not be directly relevant to satellite video tracking, necessitating task-specific adaptations. To address this, we propose the **C**ross-**S**patial **F**eature **M**apper (**CSFM**) module, which transforms model features through cross-spatial feature mapping and additional feature mapping parameters. For the i-th Transformer Block layer, the output feature is denoted as $\boldsymbol{X}_i \in \mathbb{R}^{N\times D}$, where $0 < i \leq 12$. The CSFM concatenates and integrates the output features of all Transformer Block layers, represented as $\boldsymbol{X}_C$ =

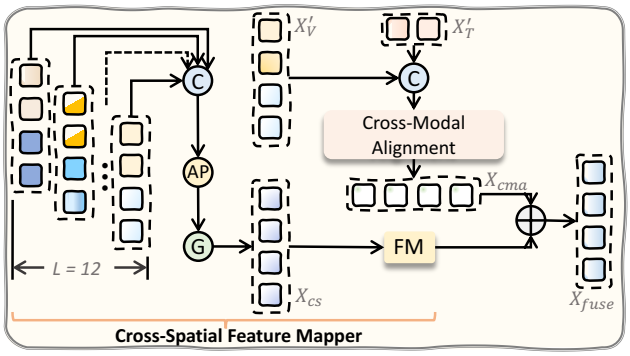

Figure 4: Detailed structure of CSFM and its relationship with CMA. "AP" is the average pooling operation, and "FM" is the feature mapper.

$Concat(\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_{12})$. The mathematical representation of CSFM can be defined as follows:

$$\boldsymbol{X}_{cs} = GELU(AP(\boldsymbol{X}_C)), \tag{4}$$

$$\boldsymbol{X}_{csfm} = FM(\boldsymbol{X}_{cs}) = \boldsymbol{X}_{cs} \cdot \boldsymbol{W}_{fm} + \boldsymbol{b}_{fm}. \tag{5}$$

Here, $AP(\cdot)$ denotes the average pooling operation, and $\boldsymbol{W}_{fm}, \boldsymbol{b}_{fm} \in \mathbb{R}^{1\times N}$ are feature mapping parameters, with $\boldsymbol{W}_{fm}$ initialized to 1 and $\boldsymbol{b}_{fm}$ initialized to 0. This operation effectively adjusts cross-spatial features to adapt to the complex scenes in satellite videos, thereby enhancing tracking performance. Additionally, we use the **C**ross-**M**odal **A**lignment (**CMA**) module to align and fuse the visual features $\boldsymbol{X}'_V$ with the textual features $\boldsymbol{X}'_T$. The resulting $\boldsymbol{X}_{cma} \in \mathbb{R}^{N\times D}$ is combined with $\boldsymbol{X}_{csfm}$ to improve tracking stability further, producing the fused feature $\boldsymbol{X}_{fuse}$. Notably, the overall structure of CMA is similar to that of DSA.

Loss Function

In the training phase, we construct historical trajectory text labels for each frame, which are represented by the coordinates of the target. The coordinate format is $[x, y, w, h]$, where x and y denote the position of the upper-left corner of the bounding box, and w and h represent the width and height of the bounding box, respectively. This sequence captures the temporal motion trajectory of the target between frames and is denoted by $\boldsymbol{T}$ (see **Section 3.1**). For the text label of the current frame, we use $\boldsymbol{T}_B$ (see **Section 3.1**). After processing these two text labels through the embedding mechanism, they are input into the LLM, generating historical trajectory label $\boldsymbol{X}_T$ and true text label $\boldsymbol{X}_{T_B}$. We employ the Mean Squared Error (MSE) loss function to constrain the generation of $\boldsymbol{X}_T$ to be consistent with $\boldsymbol{X}_{T_B}$, ensuring that the LLM produces reliable trajectory predictions. This loss function is referred to as the trajectory constraint loss $\boldsymbol{\mathcal{L}}_{mse}$, defined as follows:

$$\boldsymbol{\mathcal{L}}_{mse} = MSE(\boldsymbol{X}_T, \boldsymbol{X}_{T_B}) = \frac{1}{N_T \times D} \sum_{i=1}^{N_T \times D} (\hat{y}_i - y_i)^2, \tag{6}$$

where $\hat{y}_i$ denotes each feature value in $\boldsymbol{X}_T$, and y_i denotes the corresponding feature value in $\boldsymbol{X}_{T_B}$. Through this trajectory constraint, we ensure that the LLM effectively captures the target's spatiotemporal dependencies and motion patterns. Additionally, we use the same loss function $\boldsymbol{\mathcal{L}}_{ce}$ as (Chen et al. 2023) to supervise the tracking information. Therefore, the total loss $\boldsymbol{\mathcal{L}}_{total}$ is defined as follows:

$$\boldsymbol{\mathcal{L}}_{total} = \boldsymbol{\mathcal{L}}_{mse} + \boldsymbol{\mathcal{L}}_{ce}. \tag{7}$$

Here, $\boldsymbol{\mathcal{L}}_{ce}$ represents the cross-entropy loss used for conditioning target labels on previous sequence frames and the input video frame.

Experiments

Experimental Setup and Protocols

Model Settings. In the text branch of the proposed HTTrack, we utilize FLAN-T5-Base (Chung et al. 2024) as the LLM. The maximum number of text tokens corresponding to the history track and the target label is configured to be 128. In the visual branch, we follow the settings of (Chen et al. 2023) and use it to initialize the weights. Both the search and template images are set to be four times the area of the target object and are resized to 256×256 pixels.

Training and Tracking Strategies. Our tracker, HTTrack, is trained on two NVIDIA RTX 4090 GPUs with a global batch size of 16, and each epoch consists of 6×10^4 sample pairs. We employ data augmentation techniques, including horizontal random flipping and brightness jittering. The AdamW optimizer (Loshchilov and Hutter 2017) is used to optimize the network. The model fine-tuning requires 60 epochs, with an initial learning rate set at 0.0008, which is reduced by 10 after 48 epochs. The entire training process takes approximately 20 hours to converge. Notably, during training, the historical trajectory and target label text are jointly inputted into the text branch to learn and optimize task-related language information. For the visual branch, (Luo et al. 2023) is incorporated into the Transformer Block while freezing other parameters to adapt from the initialization parameters of (Chen et al. 2023) effectively. Note that when training the proposed HTTrack with additional annotated data from the SV248S dataset, we ensure that the video sequences in the training data do not overlap with those in the SV248S test set. This precaution is taken to avoid any potential performance bias due to data overlap. Furthermore, to ensure a fair comparison, all TF-based trackers are required to be retrained in the same SV248S training set.

During the tracking phase, only the historical trajectory is fed into the text branch and updated after each frame based on the predicted target, enabling adaptation to complex scenarios. The tracker runs in real time (greater than 30fps) on an NVIDIA 4090 GPU.

Comparative Experiments

We conduct comparative experiments on three mainstream satellite video tracking datasets, SV248S, SatSOT, and VISO (Yin et al. 2021) (the more detailed introduction to the **VISO** dataset and additional experimental comparisons are provided in the Appendix.), to evaluate the performance of HTTrack against state-of-the-art methods.

SV248S and SatSOT Datasets. The compared methods include eight CF-based trackers (e.g., CFME, ECO) and twenty-two DL-based trackers (e.g., SVLPNet, SeqTrack,

Method	Framework	Year	Features	SV248S			SatSOT	
				ENUS(↑)	Succ.(↑)	Prec.(↑)	Succ.(↑)	Prec.(↑)
MOSSE(Bolme et al. 2010)	CF-based	2010	GF	0.091	0.078	0.166	0.269	0.242
CSK(Henriques et al. 2012)	CF-based	2012	GF	-	0.050	0.089	0.247	0.237
KCF (Henriques et al. 2014)	CF-based	2014	HOG	0.253	0.417	0.736	0.393	0.521
Staple(Bertinetto et al. 2016)	CF-based	2016	HOG	-	0.147	0.343	0.382	0.462
ECO (Danelljan et al. 2017)	CF-based	2017	HOG+CN+CF	0.236	0.410	0.731	0.387	0.549
CFME (Xuan et al. 2019)	CF-based	2019	HOG	0.154	0.284	0.468	0.428	0.555
CPKF(Li, Bian, and Chen 2022)	CF-based	2022	HOG+CN+CF	-	-	-	0.468	-
DOCPF(Li et al. 2024b)	CF-based	2024	HOG+CN+CF	-	-	-	0.485	-
ATOM(Danelljan et al. 2019)	DL-based	2019	CF	0.284	0.363	0.626	0.424	0.528
SiamMask(Wang et al. 2019)	DL-based	2019	CF	0.147	0.221	0.565	0.398	0.552
DiMP (Bhat et al. 2019)	DL-based	2019	CF	0.212	0.357	0.665	0.426	0.528
SiamRPN++ (Li et al. 2019)	DL-based	2019	CF	0.213	0.335	0.663	0.423	0.537
SiamFC++(Xu et al. 2020)	DL-based	2020	CF	0.070	0.134	0.528	0.345	0.448
Ocean (Zhang et al. 2020b)	DL-based	2020	CF	0.084	0.150	0.414	-	-
SiamCAR(Guo et al. 2020)	DL-based	2020	CF	0.250	0.448	0.701	0.446	0.564
TransT (Chen et al. 2021)	DL-based	2021	CF+TF	0.192	0.267	0.559	0.388	0.496
SiamGAT (Guo et al. 2021)	DL-based	2021	CF	0.227	0.376	0.688	-	-
STARK (Yan et al. 2021)	DL-based	2021	CF+TF	0.220	0.363	0.624	0.345	0.404
OSTrack (Ye et al. 2022)	DL-based	2022	TF	0.244	0.399	0.659	0.359	0.431
SimTrack (Chen et al. 2022)	DL-based	2022	TF	0.216	0.350	0.600	0.333	0.409
ARTrack (Wei et al. 2023)	DL-based	2023	TF	0.260	0.437	0.739	-	-
SeqTrack (Chen et al. 2023)	DL-based	2023	TF	0.272	0.443	0.705	0.427	0.512
TATrans (Lai et al. 2023)	DL-based	2023	CF+TF	-	-	-	0.456	0.576
ARTrackV2 (Bai et al. 2024)	DL-based	2024	TF	0.261	0.447	0.755	-	-
RRT (Yang et al. 2024)	DL-based	2024	CF	-	0.457	0.710	-	-
SVLPNet (Wang et al. 2024a)	DL-based	2024	CF+TF	0.279	0.460	0.783	0.465	0.584
ARTrackV2 (Bai et al. 2024)	DL-based	2024	TF	0.231	0.397	0.731	-	-
ODTrack (Zheng et al. 2024)	DL-based	2024	TF	0.211	0.376	0.736	-	-
SGLATrack (Xue et al. 2025)	DL-based	2025	TF	0.256	0.420	0.712	0.416	0.513
ORTarck (Wu et al. 2025)	DL-based	2025	TF	0.251	0.414	0.706	0.429	0.534
HTTrack (Ours)	DL-based	Ours	TF	**0.313**	**0.515**	**0.809**	**0.529**	**0.650**

Table 1: Overall performance on SV248S and SatSOT. Highlighting the top three highest scores in **red**, **green**, and **blue**, respectively. Notably, the abbreviations used denote different feature extraction methods: HOG for Histogram of Oriented Gradient, CN for Color Name, CF for Convolutional Feature, and TF for Transformer.

ARTrack), as shown in Table 1. The results demonstrate that HTTrack outperforms the CF-based tracker CFME across three evaluation metrics on the SV248S dataset, with 15.9%, 23.1%, and 34.1% improvements in ENUS, SR, and PR, respectively. On the SatSOT dataset, HTTrack also shows a notable increase of 10.1% and 9.5% in SR and PR, respectively. Compared to the latest state-of-the-art satellite video tracker SVLPNet, which uses multiple updatable visual templates, HTTrack significantly improves both datasets. This highlights the effectiveness of leveraging historical trajectories augmented by LLMs to enhance tracking performance. These results suggest that incorporating historical trajectory information, as processed by LLMs, provides valuable guidance to the tracker, leading to more accurate and robust tracking in satellite video scenarios.

We use attribute radar charts to visually compare HTTrack and other trackers based on key metrics from the SV248S dataset, which includes 10 attributes: *Short-Term Occlusion (STO), Long-Term Occlusion (LTO), Dense Similarity (DS), Illumination Variation (IV), Background Change (BCH), Slow Motion (SM), Natural Disturbance (ND), Continuous Occlusion (CO), Background Cluster (BCL), and In-Plane Rotation (IPR)*. Figure 5 clearly illustrates that HTTrack consistently achieves higher scores across most attributes than other methods. This superior performance underscores HTTrack's strong adaptability across different scenarios and conditions. By effectively integrating historical trajectory information with visual features, HTTrack can handle diverse tracking challenges, such as scale variations, occlusions, and complex backgrounds, more efficiently than its counterparts. This adaptability reinforces HTTrack as a robust and reliable solution for satellite video tracking.

Ablation Experiments

Impact of Different Components. To evaluate the contribution of key components in the proposed HTTrack, we conduct a comprehensive ablation study on different variants of HTTrack and the foundation model. The results are presented in Table 2. Starting from the foundation model, we first introduce the **Adapter** (the details are provided in Ap-

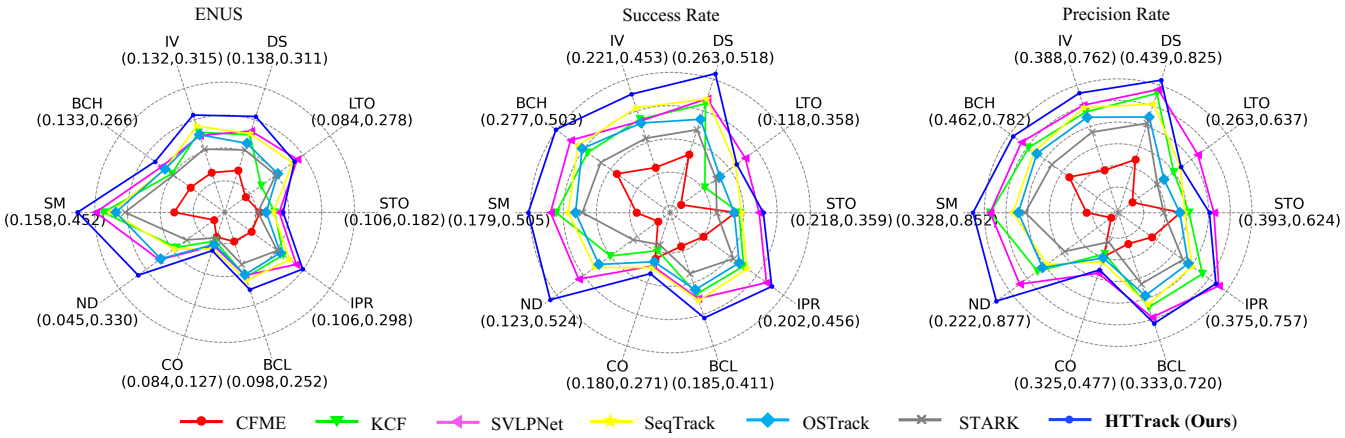

Figure 5: Radar charts of ten different attributes on the SV248S test set under various evaluation metrics.

Options	SR(↑)	PR(↑)
Foundation model (SeqTrack)	44.3	70.5
+ Adapter	46.9$_{+2.6}$	75.3$_{+4.8}$
+ Text Branch	48.8$_{+4.5}$	77.1$_{+6.6}$
+ DSA	49.5$_{+5.2}$	77.8$_{+7.3}$
+ CMA	49.9$_{+5.6}$	79.0$_{+8.5}$
+ CSFM	50.9$_{+6.6}$	79.2$_{+8.7}$
+ Similarity	**51.5**$_{+7.2}$	**80.9**$_{+10.4}$

Table 2: Quantifying the contribution of different components in the SV248S test set.

pendix), which significantly improves SR and PR by 2.9% and 4.8%, respectively. This indicates that including the Adapter helps maintain the stability of the original knowledge in the foundation model, preventing the loss of previously learned useful information when training on new tasks. Next, we ablate the **Text Branch**. SR and PR improve to 48.1% and 77.1% upon adding the text branch, respectively. This notable enhancement underscores the critical role of historical trajectories in boosting the tracker's performance. To further enhance performance, we employ **DSA** and **CMA** to align historical trajectory information with visual features. We obtain complementary information from both modalities to handle challenging scenarios better and improve performance. Additionally, we introduce the **CSFM** module to fully leverage spatial feature information to remap feature embeddings across all Transformer Block layers. This adaptation aims to better suit the model to the complex scenes in satellite videos. As expected, this leads to a significant increase in performance, with SR and PR reaching 50.9% and 79.2%, respectively. Finally, we incorporate the current frame's label text in the text branch to constrain the historical trajectory, generating features more relevant to the current frame's target. This results in the best performance, with SR and PR achieving 51.5% and 80.9%, respectively. These results effectively validate the effectiveness of the proposed HTTrack in enhancing tracking performance.

Exploration of Different LLMs. To assess the effect of various LLMs on HTTrack's performance, we conduct an ablation study comparing several popular LLMs, including CLIP$_{Text}$, DeBERTa, T5, Qwen2.5, and FLAN-T5. The outcomes of these comparisons are shown in Table 3. FLAN-T5 demonstrated the highest performance among the evaluated models, indicating its exceptional capacity to efficiently interpret and leverage historical trajectory text information to improve the tracker. Employing FLAN-T5 as the language model in HTTrack yields the most precise tracking outcomes, highlighting its efficacy in handling diverse and complex textual data in satellite video scenarios.

LLM	SR(↑)	PR(↑)
HTTrack (CLIP$_{Text}$ (Radford et al. 2021))	49.4	77.8
HTTrack (DeBERTa (He et al. 2020))	50.5	78.5
HTTrack (T5 (Raffel et al. 2020))	51.1	79.7
HTTrack (Qwen2.5 (Yang et al. 2025))	50.2	77.1
HTTrack (FLAN-T5)	**51.5**	**80.9**

Table 3: Compare the effects of different LLMs in HTTrack on the SV248S test set.

More ablation experiments (Impact of the number of historical trajectories and different DSA module variants) and visualization analyses are provided in Appendix.

Conclusion

In this work, we propose HTTrack, a novel satellite video tracking method that integrates historical trajectories and visual templates to enhance tracking performance. Unlike traditional methods relying solely on visual templates, HTTrack employs a hybrid approach that leverages historical trajectory information and visual features. We introduce several innovative components, including the DSA, CMA, and CSFM modules, to address the challenges of target appearance changes, motion patterns, and occlusions in satellite video tracking. Our extensive experiments on the SV248S, SatSOT, and VISO datasets demonstrate that HTTrack significantly outperforms state-of-the-art tracking methods, including CF-based and DL-based trackers.

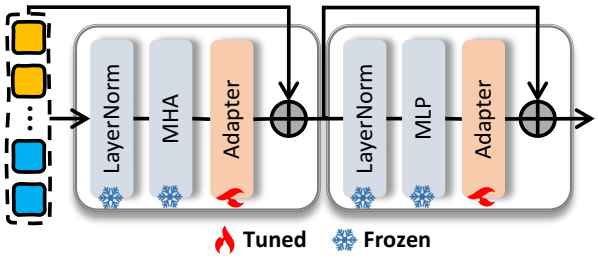

Figure 6: Structural details of the Adapter.

Number	1	3	5	10
SR(↑)	50.6	51.2	**51.5**	48.9
PR(↑)	79.1	80.3	**80.9**	78.4

Table 4: The performance of the proposed HTTrack on the SV248S test set when setting different numbers of historical trajectories.

Supplemental Material

This section provides supplementary materials, offering additional detailed information to complement the main paper and further experimental analysis. The content in this section is organized as follows.

- **Insert Adapter in Transformer Block**
- **More Ablation Experiments**
- **Visualization and Analysis**
- **Comparative Experiments on the VISO Dataset**
- **Qualitative Comparison**

Insert Adapter in Transformer Block

Figure 6 illustrates the insertion of Adapters into a Transformer Block. Both Adapters share the same structure but serve distinct roles within the block. The first Adapter is placed in the Multi-Head Attention (MHA) section to fine-tune attention outputs and ensure that the MHA effectively captures relevant information. At the same time, the second is positioned in the Multi-Layer Perceptron (MLP) section to refine the transformations applied by the MLP, enabling better adaptation of the model to the specific task. In both sections, the Adapter is the only fine-tuned component. In contrast, the rest of the components (LayerNorm, MHA, MLP) are frozen. This approach allows the model to leverage pre-trained knowledge while still adapting to the new task through fine-tuning the Adapters.

More Ablation Experiments

Impact of the Number of Historical Trajectories. To further investigate the impact of the number of historical trajectories on the performance of HTTrack, we conduct an ablation study with varying numbers of historical trajectories. Specifically, we set the number of historical trajectories to 1, 3, 5, and 10. Table 4 indicates that the tracking performance improves as the number of historical trajectories increases, with a peak performance observed at five historical trajectories. This suggests that using five historical trajectories provides an optimal balance between capturing sufficient past information and maintaining computational efficiency. Beyond this point, additional trajectories do not contribute significantly to performance improvements, possibly due to diminishing returns in the relevant historical context.

Method	SR(↑)	PR(↑)
Only textual feature tokens	49.9	78.7
Remove scaling factors	50.6	80.1
HTTrack	**51.5**	**80.9**

Table 5: The impact of different DSA module variants on the performance of HTTrack on the SV248S dataset.

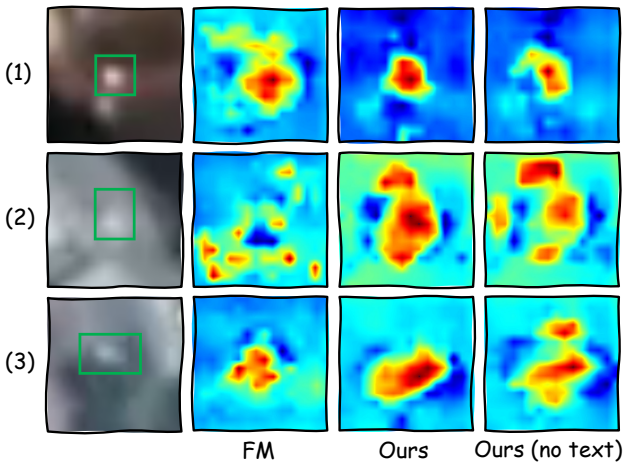

Figure 7: Visualization of response maps on representative scenarios from the SV248S test set. The green box represents the ground truth target location.

The Impact of Different DSA Module Variants. We conduct an ablation study to explore the impact of different variants of the proposed DSA module on the performance of HTTrack. Specifically, we examine two variations: using only textual feature tokens and removing the scaling factors. Table 5 shows that performance significantly decreases when either variant is applied, indicating the importance of textual-visual alignment and dynamic scaling in achieving optimal tracking accuracy. The findings indicate that the DSA module, incorporating feature integration and scaling, is essential for augmenting HTTrack's capacity to manage intricate tracking situations.

Visualization and Analysis

We conduct a visual comparison of representative scenarios from the SV248S test set, including dense similarity, background change, and occlusions, to evaluate the foundation model (FM), the proposed HTTrack, and its variant without textual information, as shown in Figure 7. In the first scenario with dense similarity, HTTrack achieves a highly

	KCF	ECO	MDNet	RT-MDNet	TransT	SiamGAT	CFME	SiamRPN++	DiMP	PrDiMP	SeqTrack	**HTTrack**
SR(↑)	0.185	0.242	0.228	0.226	0.098	0.152	0.217	0.232	0.203	0.181	0.260	**0.326**
PR(↑)	0.462	0.607	0.622	0.646	0.384	0.490	0.547	0.522	0.583	0.450	0.567	**0.681**

Table 6: Overall performance on VISO dataset. The top three highest scores are highlighted in **red**, **green**, and **blue**.

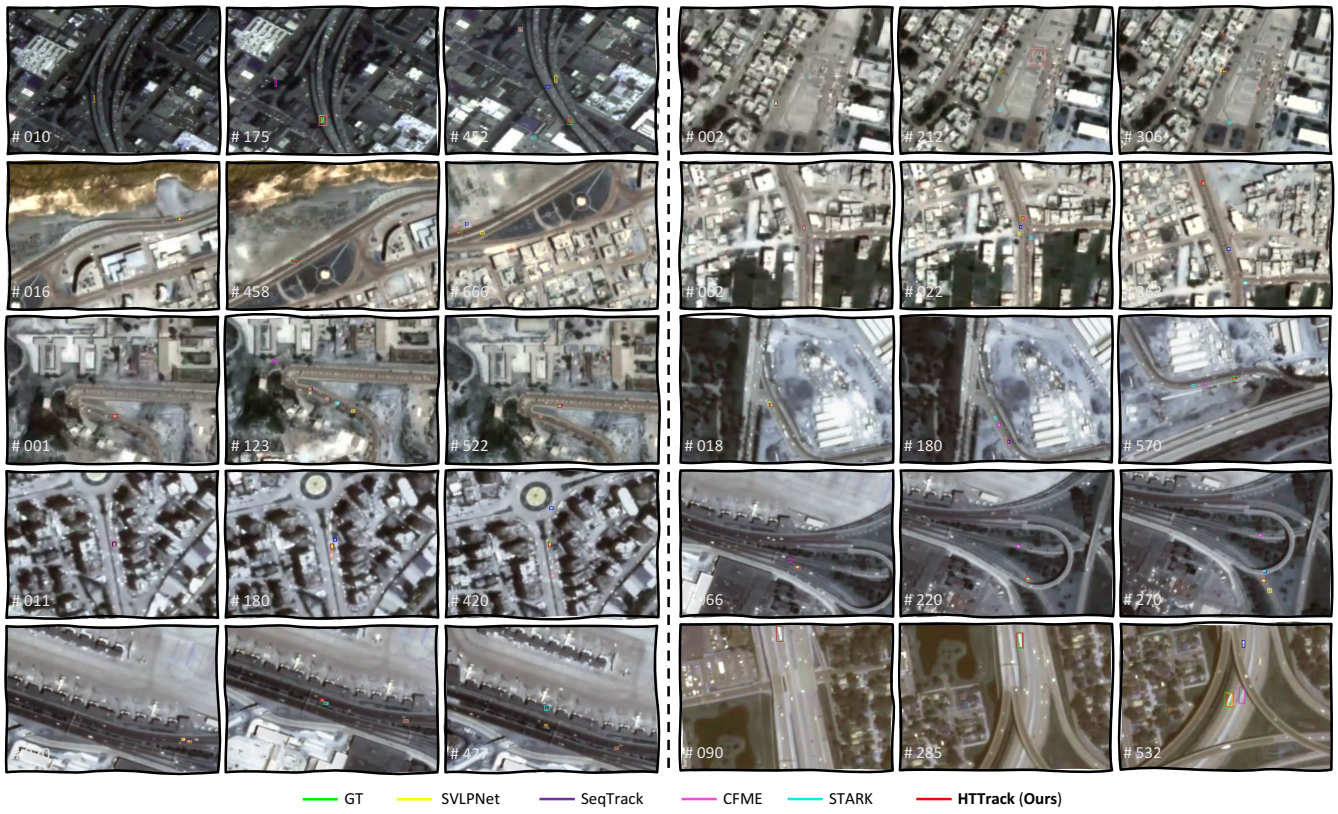

Figure 8: Qualitative results from some state-of-the-art trackers on challenging video sequences in the SV248S dataset. It is recommended that you zoom in to view the tracking results.

focused response on the target, effectively suppressing interference, whereas the FM and the text-free variant show scattered activations. In the second scenario, where the target blends into the background, HTTrack produces concentrated responses, successfully distinguishing the target, while the other models suffer from background interference. In the third scenario involving occlusions, HTTrack maintains a clear focus on the target, demonstrating strong robustness, while the baseline and text-free variant exhibit noisy or partially accurate responses. These results highlight that the integration of textual information in HTTrack significantly enhances target localization accuracy and robustness under complex conditions.

Comparative Experiments on the VISO

In our experiments on the VISO dataset, HTTrack significantly outperforms other state-of-the-art trackers, as illustrated in Table 6. HTTrack achieves a Success Rate of 32.6% and a Precision Rate of 68.1%, surpassing the second-best tracker, ECO, by 8.4% in Success Rate and 7.4% in Precision Rate. These results underscore the significant advantage of HTTrack, particularly in leveraging historical trajectories to manage the complex challenges of satellite video data. Using historical information, HTTrack enhances its ability to accurately track small, fast-moving objects within high-resolution satellite imagery, demonstrating superior robustness and precision in dynamic and challenging scenarios.

Qualitative Comparison

In the qualitative comparison, HTTrack demonstrates superior tracking performance compared to other methods, such as SVLPNet, SeqTrack, CFME, and STARK, as visualized in Figure 8. The proposed HTTrack stands out because it integrates historical trajectory information with visual features, enabling it to maintain higher accuracy and stability across various challenging scenarios. This advantage is attributed to HTTrack's ability to effectively leverage temporal and spatial information, ensuring robust tracking even with significant target scale variations, orientation changes, and complex backgrounds.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China(No.62576264), Project supported by the National Science and Technology Major Project of the Ministry of Science and Technology of China (No.2025ZD0551500, No.2025ZD0551502), the Key Project of National Natural Science Foundation of China (62431020,62231027), the Joint Fund Project of National Natural Science Foundation of China (No.U22B2054), the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) (No.B07048), the Postdoctoral Fellowship Program of China Postdoctoral Science Foundation (CPSF) (No.GZC20232033), the Program for Cheung Kong Scholars and Innovative Research Team in University (No.IRT 15R53), the Key Scientific Technological Innovation Research Project by Ministry of Education and the National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University (No.HMHAI-202404, No. HMHAI-202405).

References

Bai, Y.; Zhao, Z.; Gong, Y.; and Wei, X. 2024. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proc. of CVPR*.

Bertinetto, L.; Valmadre, J.; Golodetz, S.; Miksik, O.; and Torr, P. H. 2016. Staple: Complementary learners for real-time tracking. In *Proc. of CVPR*.

Bhat, G.; Danelljan, M.; Gool, L. V.; and Timofte, R. 2019. Learning discriminative model prediction for tracking. In *Proc. of ICCV*.

Bolme, D. S.; Beveridge, J. R.; Draper, B. A.; and Lui, Y. M. 2010. Visual object tracking using adaptive correlation filters. In *Proc. of CVPR*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Proc. of NeurIPS*.

Chen, B.; Li, P.; Bai, L.; Qiao, L.; Shen, Q.; Li, B.; Gan, W.; Wu, W.; and Ouyang, W. 2022. Backbone is all your need: A simplified architecture for visual object tracking. In *Proc. of ECCV*.

Chen, P.; Wang, L.; Guo, L.; Liu, X.; Zhang, X.; Jiao, L.; and Liu, F. 2024. Satellite Videos Object Tracking Based on Enhanced Correlation Filter with Motion Prediction Network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Chen, X.; Peng, H.; Wang, D.; Lu, H.; and Hu, H. 2023. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking. In *Proc. of CVPR*.

Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; and Lu, H. 2021. Transformer tracking. In *Proc. of CVPR*.

Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; and Ji, R. 2020. Siamese box adaptive network for visual tracking. In *Proc. of CVPR*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*.

Danelljan, M.; Bhat, G.; Khan, F. S.; and Felsberg, M. 2019. Atom: Accurate tracking by overlap maximization. In *Proc. of CVPR*.

Danelljan, M.; Bhat, G.; Shahbaz Khan, F.; and Felsberg, M. 2017. Eco: Efficient convolution operators for tracking. In *Proc. of CVPR*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Guo, D.; Shao, Y.; Cui, Y.; Wang, Z.; Zhang, L.; and Shen, C. 2021. Graph attention tracking. In *Proc. of CVPR*.

Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; and Chen, S. 2020. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In *Proc. of CVPR*.

He, A.; Luo, C.; Tian, X.; and Zeng, W. 2018. A twofold siamese network for real-time object tracking. In *Proc. of CVPR*.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

He, Q.; Zeng, J.; Huang, W.; Chen, L.; Xiao, J.; He, Q.; Zhou, X.; Liang, J.; and Xiao, Y. 2024. Can Large Language Models Understand Real-World Complex Instructions? In *Proc. of AAAI*.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2012. Exploiting the circulant structure of tracking-by-detection with kernels. In *Proc. of ECCV*.

Henriques, J. F.; Caseiro, R.; Martins, P.; and Batista, J. 2014. High-speed tracking with kernelized correlation filters. *IEEE transactions on pattern analysis and machine intelligence*.

Huang, Z.; Jiao, L.; Zhang, J.; Liu, X.; Liu, F.; Zhang, X.; Li, L.; and Chen, P. 2024. A Graph Association Motion-aware Tracker for Tiny Object in Satellite Videos. *IEEE Transactions on Circuits and Systems for Video Technology*.

Jiao, L.; Wang, D.; Bai, Y.; Chen, P.; and Liu, F. 2021a. Deep learning in visual tracking: A review. *IEEE transactions on neural networks and learning systems*.

Jiao, L.; Zhang, R.; Liu, F.; Yang, S.; Hou, B.; Li, L.; and Tang, X. 2021b. New generation deep learning for video object detection: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Jiao, L.; Zhang, X.; Liu, X.; Liu, F.; Yang, S.; Ma, W.; Li, L.; Chen, P.; Feng, Z.; Guo, Y.; et al. 2023. Transformer meets remote sensing video detection and tracking: A comprehensive survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Lai, P.; Zhang, M.; Cheng, G.; Li, S.; Huang, X.; and Han, J. 2023. Target-aware transformer for satellite video object tracking. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; and Yan, J. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proc. of CVPR*.

Li, B.; Yan, J.; Wu, W.; Zhu, Z.; and Hu, X. 2018. High performance visual tracking with siamese region proposal network. In *Proc. of CVPR*.

Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; and Qiao, Y. 2023a. VideoChat: Chat-Centric Video Understanding. *arXiv preprint arXiv:2305.06355*.

Li, X.; Jiao, L.; Zhu, H.; Huang, Z.; Liu, F.; Li, L.; Chen, P.; and Yang, S. 2023b. A complex-former tracker with dynamic polar spatio-temporal encoding. *IEEE Transactions on Neural Networks and Learning Systems*.

Li, Y.; Bian, C.; and Chen, H. 2022. Object tracking in satellite videos: Correlation particle filter tracking method with motion estimation by Kalman filter. *IEEE Transactions on Geoscience and Remote Sensing*.

Li, Y.; Hu, B.; Shi, H.; Wang, W.; Wang, L.; and Zhang, M. 2024a. VisionGraph: Leveraging Large Multimodal Models for Graph Theory Problems in Visual Context. *arXiv preprint arXiv:2405.04950*.

Li, Y.; Wang, N.; Li, W.; Li, X.; and Rao, M. 2024b. Object tracking in satellite videos with distractor–occlusion-aware correlation particle filters. *IEEE Transactions on Geoscience and Remote Sensing*.

Liang, Z.; and Shen, J. 2019. Local semantic siamese networks for fast tracking. *IEEE Transactions on Image Processing*.

Liu, F.; Wang, J.; Jiao, L.; Zhang, J.; Wang, H.; Li, S.; Li, L.; Chen, P.; Liu, X.; Ma, W.; et al. 2025. Remote Sensing Video Tracking: Current Status, Challenges and Future. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, G.; Huang, M.; Zhou, Y.; Sun, X.; Jiang, G.; Wang, Z.; and Ji, R. 2023. Towards efficient visual adaption via structural re-parameterization. *arXiv preprint arXiv:2302.08106*.

Pang, R.; Gao, F.; Zhang, P.; Li, X.; and Zhai, Y. 2023. Aircraft Tracking Based on an Antidrift Multifilter Tracker in Satellite Video Data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*.

Shang, R.; Xie, K.; Okoth, M. A.; and Jiao, L. 2019. SAR image change detection based on mean shift preclassification and fuzzy C-means. In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*.

Shao, J.; Du, B.; Wu, C.; Gong, M.; and Liu, T. 2021. Hrsiam: High-resolution siamese network, towards spaceborne satellite video tracking. *IEEE Transactions on Image Processing*.

Shao, Y.; Guo, D.; Cui, Y.; Wang, Z.; Zhang, L.; and Zhang, J. 2024. Graph Attention Network for Context-Aware Visual Tracking. *IEEE Transactions on Neural Networks and Learning Systems*.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Proc. of NeurIPS*.

Sun, T.; Zhang, X.; He, Z.; Li, P.; Cheng, Q.; Liu, X.; Yan, H.; Shao, Y.; Tang, Q.; Zhang, S.; et al. 2024. MOSS: An Open Conversational Large Language Model. *Machine Intelligence Research*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, J.; Liu, F.; Jiao, L.; Gao, Y.; Wang, H.; Li, L.; Chen, P.; Liu, X.; and Li, S. 2024a. Satellite Video Object Tracking based on Location Prompts. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, J.; Liu, F.; Jiao, L.; Gao, Y.; Wang, H.; Li, S.; Li, L.; Chen, P.; and Liu, X. 2024b. Visual and Language Collaborative Learning for RGBT Object Tracking. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, J.; Liu, F.; Jiao, L.; Wang, H.; Li, S.; Li, L.; Chen, P.; and Liu, X. 2024c. Multi-modal visual tracking based on textual generation. *Information Fusion*.

Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; and Torr, P. H. 2019. Fast online object tracking and segmentation: A unifying approach. In *Proc. of CVPR*.

Wang, W.; Chen, Z.; Chen, X.; Wu, J.; Zhu, X.; Zeng, G.; Luo, P.; Lu, T.; Zhou, J.; Qiao, Y.; et al. 2024d. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Proc. of NeurIPS*.

Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; and Gong, Y. 2023. Autoregressive visual tracking. In *Proc. of CVPR*.

Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; and Duan, N. 2023a. Visual chatgpt: Talking, drawing and editing with visual foundation models. *arXiv preprint arXiv:2303.04671*.

Wu, M.; Gu, J.; Shen, Y.; Lin, M.; Chen, C.; and Sun, X. 2023b. End-to-end zero-shot hoi detection via vision and language knowledge distillation. In *Proc. of AAAI*.

Wu, Y.; Wang, X.; Yang, X.; Liu, M.; Zeng, D.; Ye, H.; and Li, S. 2025. Learning Occlusion-Robust Vision Transformers for Real-Time UAV Tracking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; and Yu, G. 2020. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *Proc. of AAAI*.