SSTtrack: A Unified Hyperspectral Video Tracking Framework via Modeling Spectral-Spatial-Temporal Conditions

Yuzeng Chen [a], Qiangqiang Yuan [a, b, *], Yuqi Tang [c], Yi Xiao [a], Jiang He [a], Te Han [c], Zhenqi Liu [d], and Liangpei Zhang [e]

[a] *School of Geodesy and Geomatics, Wuhan University, Wuhan, Hubei 430079, China*
[b] *Hubei Luojia Laboratory, Wuhan, Hubei 430079, China*
[c] *School of Geosciences and Info-Physics, Central South University, Changsha, China*
[d] *College of Artificial Intelligence, Southwest University, Chongqing, 400715, China*
[e] *State Key Laboratory of Information Engineering, Survey Mapping and Remote Sensing, Wuhan University, Wuhan, China*
** Corresponding author*

ARTICLEINFO

Keywords:
Hyperspectral video
Spectral awareness
Temporal awareness
Prompt learning
Multi-modal tracking

ABSTRACT

Hyperspectral video contains rich spectral, spatial, and temporal conditions that are crucial for capturing complex object variations and overcoming the inherent limitations (e.g., multi-device imaging, modality alignment, and finite spectral bands) of regular RGB and multi-modal video tracking. However, existing hyperspectral tracking methods frequently encounter issues including data anxiety, band gap, huge volume, and weakness of the temporal condition embedded in video sequences, which result in unsatisfactory tracking performance. To address these dilemmas, we propose a unified hyperspectral video tracking framework via modeling spectral-spatial-temporal conditions in an end-to-end fashion, dubbed SSTtrack. First, we design a multi-modal generation adapter (MGA) to explore the interpretability benefits of combining physical and machine models for learning the multi-modal generation and bridging the band gap. Then, we construct a spectral-spatial adapter (SSA) to dynamically transfer and interact with multiple modalities. Finally, we design a temporal condition adapter (TCA) for injecting the temporal condition to guide spectral and spatial feature representations to capture static and instantaneous object properties. SSTtrack follows the prompt learning paradigm with the addition of fewer trainable parameters (0.575M), resulting in superior performance in extensive comparisons. The code and model will be available at **https://github.com/YZCU/SSTtrack**.

1. Introduction

Visual tracking, essential for establishing the association of the object in videos, has diverse applications including augmented reality, autonomous vehicles, and industrial automation (Javed et al., 2022). While progress using RGB modality is notable, challenges persist in scenarios with similar appearances, low light, and background clutter due to limited spectral information. To conquer these challenges, multi-modal tracking has emerged via combining RGB with event (RGB-E) (Zhu et al., 2023a), thermal infrared (RGB-T) (Cao et al., 2024), and depth (RGB-D) (Xuefeng Zhu et al., 2024) data, as shown in Fig. 1. However, regular multi-modal tracking requires multiple imaging devices, posing challenges in capturing the same scene accurately, especially for distant or small objects (Chen et al., 2024a). Consequently, aligning multiple modalities has become a common procedure (Li et al., 2022), albeit potentially causing image distortion, as illustrated in Figs. 1(a) and (b). Moreover, regular multi-modal tracking typically couples only two modalities in an RGB+X pattern (e.g., RGB-E, RGB-T, and RGB-D) (Zhu et al., 2023a), yielding suboptimal results in complex scenarios. Indeed, multi-modal data comprises information from various radiated bands (e.g., RGB and thermal infrared) or different mechanisms (e.g., event and depth). Hyperspectral (HS) cameras capture material-radiated signals across distinct bands, enabling trackers to identify object materials (Xiong et al., 2020). The abundant HS bands, captured from the same viewpoint, can generate multiple modalities. Thus, extending HS video tracking to multi-modal tracking is expected to address modality alignment and spectral limitations (Chen et al., 2024a). However, effectively utilizing multi-modal information, including spectral, spatial, and temporal conditions, remains a challenge in HS video tracking.

Currently, HS video tracking encounters several dilemmas. First, data anxiety refers to the scarcity of HS videos, which hinders the direct training of robust tracking models (Liu et al., 2022). Second, the band gap between RGB and HS modalities makes it difficult to dynamically utilize rich spectral and spatial conditions (Chen et al., 2024b). Third, the huge volume caused by dense bands slows down tracking speed (Sun et al., 2023). At last, the temporal condition is far from being mined, albeit invaluable for robust tracking (Chen et al., 2024a). To address the data anxiety, one conceivable idea is to leverage large-scale RGB tracking datasets like TrackingNet (Muller et al., 2018), LaSOT (Fan et al., 2019), and COCO (Lin et al., 2014) for initial training of foundation models followed by performing full fine-tuning on HS training sets. However, fine-tuning poses risks of overfitting due to the scarcity of HS video datasets. Additionally, full fine-tuning is time-consuming and resource-intensive compared to promising prompt-tuning (Zhu et al., 2023a). To bridge the band gap, certain methods convert HS images into a three-channel representation, such as SiamHT (Tang et al., 2023), BAHT (Tang et al., 2022b), BS-SiamRPN (Wang et al., 2022), (Su et al., 2022), and (Zhang et al., 2022b). Nonetheless, this conversion inevitably results in the loss and distortion of the vital spectral condition (Chen et al., 2024b). Others focus on leveraging rich spectral conditions, yielding competitive results, such as DSP-Net (Zhu et al., 2023b), MHT (Xiong et al., 2020), SMT (Chen et al., 2023a), SiamBAG (Li et al., 2023a), CBFF-Net (Gao et al., 2023b), Trans-HST (Gao et al., 2023a), and TFTN (Zhao et al., 2022). Regarding the large volume, decision-level fusion methods are common but can limit efficiency

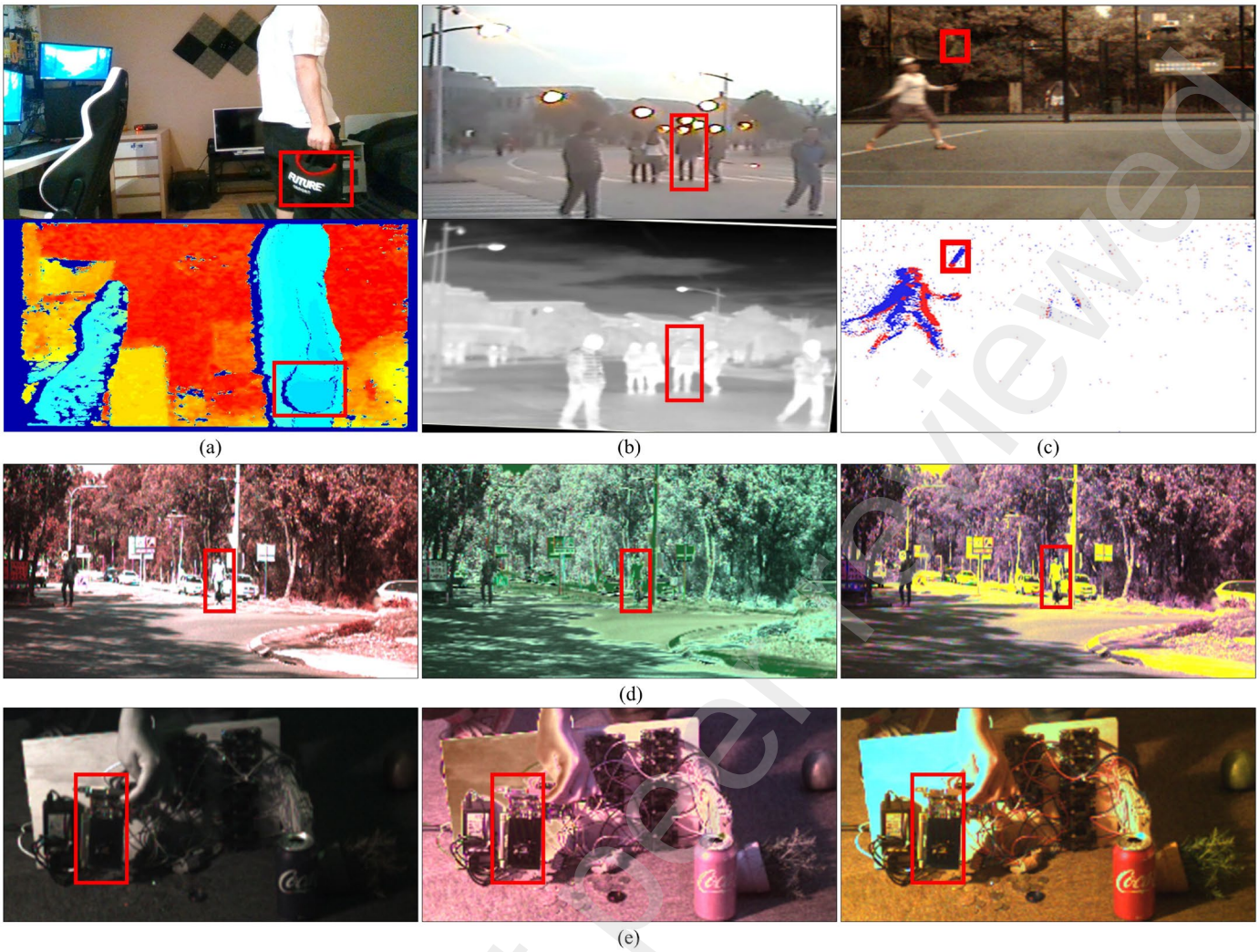

Fig. 1. Sample of multi-modal data. (a) RGB and depth modalities. (b) RGB and thermal infrared modalities. (c) RGB and event modalities. (d) and (e) Two samples of multi-modal data generated by the hyperspectral modality.

by producing multiple weak results in each inference session, such as SST-Net (Li et al., 2021), BRRF-Net (Ouyang et al., 2022), SEE-Net (Li et al., 2023b), and BAE-Net (Li et al., 2020b). In addition, previous studies have mainly focused on spatial-spectral conditions and struggled to dynamically fuse visual feature prompts across modalities. Few methods have integrated the temporal condition contained in videos for HS tracking, such as SPIRIT (Chen et al., 2024b), SENSE (Chen et al., 2024a), and SST-Net (Li et al., 2021). Despite their success, these methods often necessitate the manual design of update strategies and the introduction of hyper-parameters (e.g., intervals and thresholds), which fall short of adequately exploiting the temporal condition.

Motivated by the above analysis, we propose SSTtrack, a unified HS video tracking framework via modeling spectral-spatial-temporal conditions in an end-to-end fashion. Adhering to the prompt learning paradigm, SSTtrack harnesses the potential of foundation models to address data anxiety. The framework comprises three main components: the multi-modal generation adapter (MGA), spectral-spatial adapter (SSA), and temporal condition adapter (TCA). First, we design the MGA to bridge the band gap and stimulate prior knowledge. Drawing on the interpretability advantages of combining physical and machine models, MGA achieves multi-modal generation adaptively. Next, we construct the SSA to fully exploit spectral and spatial conditions. SSA comprises a quintuple-stream architecture with an interactive modal perception module to learn prompts from other modalities. Finally, TCA is designed to inject the temporal condition for guiding spectral and spatial feature representations. Notably, the feature-level fusion strategy permeates SSTtrack, alleviating the huge volume.

Major contributions are outlined as follows:

- To fully model spectral-spatial-temporal conditions, we propose SSTtrack, a unified hyperspectral video tracking framework via modeling spectral-spatial-temporal conditions in an end-to-end fashion. SSTtrack inherits the prompt learning paradigm by adding fewer trainable parameters (0.575M), empowering the foundation model to effectively tackle data anxiety.
- A multi-modal generation adapter (MGA) is proposed to learn the multi-modal generation by capitalizing on the interpretability benefits of integrating physical and machine models. Benefiting from MGA, SSTtrack can bridge the band gap to stimulate prior knowledge of foundation models.
- A spectral-spatial adapter (SSA) is proposed to mine the spectral-spatial conditions. Within SSA, an interactive modal perception module (IMPM) facilitates the transfer and interaction of complementary features across modalities, enhancing spectral-spatial feature representations.
- A temporal condition adapter (TCA) is proposed to inject the temporal condition to guide spectral-spatial conditions. Within TCA, a spectral-spatial-temporal filtering module (SSTFM) filters spectral-spatial-temporal conditions, facilitating the

integration of static and instantaneous object properties.

We conduct extensive experiments to validate the SSTtrack framework. The remainder is organized as follows: Section II reviews related research. In Section III, we detail the proposed method. Section IV presents the results and analysis. Finally, Section V formulates the conclusions and highlights the main contributions.

2. Review on related research

2.1. Tracking via HS video

Generative and discriminative paradigms are prevalent in HS video tracking (Chen et al., 2024a). Initially, researchers concentrate on the generative paradigm, developing models to represent objects and identify similar regions. Recently, HS trackers have shifted towards the discriminative paradigm, which includes correlation filters and Siamese networks. Certain HS trackers aim to leverage full band information by correlation filter models, such as TASSCF (Tang et al., 2022a), MFI (Zhang et al., 2021), MHT (Xiong et al., 2020), and TSCFW (Hou et al., 2022). Despite the integration of hand-crafted and/or deep features, they have encountered limited success, largely due to the relative simplicity of the model (Li et al., 2023b). Siamese networks have received significant attention in the HS tracking domain. HS trackers like MMF-Net (Li et al., 2024), SiamOHOT (Sun et al., 2023), SEE-Net (Li et al., 2023b), SiamBAG (Li et al., 2023a), SiamHYPER (Liu et al., 2022), SiamCAT (Jiang et al., 2024), CBFF-Net (Gao et al., 2023b), DSP-Net (Zhu et al., 2023b), SiamHT (Tang et al., 2023), BRRF-Net (Ouyang et al., 2022), and Trans-HST (Gao et al., 2023a) have integrated Siamese networks to achieve outstanding performance, laying a solid foundation for our research. For instance, MMF-Net (Li et al., 2024) combines HS, false color, and material information to mitigate tracking drift. CBFF-Net (Gao et al., 2023b) introduces a bidirectional multiple deep feature fusion module and a cross-band group attention module to interact with multiple bands for stable performance. SiamCAT (Jiang et al., 2024) introduces a channel adaptive module and a guided learning attention module to improve effectiveness. However, previous studies have mainly focused on spatial-spectral conditions and struggled to dynamically fuse visual feature prompts across modalities. Few methods have integrated the temporal condition contained in videos for HS tracking, such as SPIRIT (Chen et al., 2024b), SST-Net (Li et al., 2021), and SENSE (Chen et al., 2024a). SPIRIT (Chen et al., 2024b) integrates an update module to assess the tracking confidence for adapting to object changes and mitigating tracking drift. SST-Net (Li et al., 2021) inherits the temporal attention with RNN-like structure to select valuable bands for deep ensemble tracking. While SENSE (Chen et al., 2024a) designs a motion-aware module comprising an awareness selector to determine the reliability of material and motion cues and a motion prediction scheme to manage abnormal states, facilitating continuous tracking. Despite their success, these methods often require manual design of update strategies and the introduction of hyper-parameters (e.g., intervals and thresholds), which falls short in adequately exploiting the temporal condition. Moreover, the aforementioned challenges of data anxiety, band gap, and huge volume inherently impede the robustness of HS video tracking. To alleviate these dilemmas, we propose SSTtrack, a unified HS video tracking framework via modeling spectral-spatial-temporal conditions in an end-to-end fashion, attaining comparable tracking results.

2.2. Multi-modal tracking via prompt learning

Recently, prompt learning has emerged as a technique that significantly enhances the performance of various language-processing tasks (Liu et al., 2023). Interestingly, this approach is gaining traction in the visual tracking community (Han et al., 2024; Zhu et al., 2023a). For instance, ProTrack (Yang et al., 2022) converts multi-modal inputs into a single modality using the prompt paradigm, achieving effective tracking capabilities learned from foundation models. ViPT (Zhu et al., 2023a) introduces prompt learning to multi-modal tracking by learning prompts relevant to each modality. OneTrack (Han et al., 2024) employs prompt-tuning techniques for RGB+X tracking tasks by pre-training the foundation model on the RGB dataset. Benefitting from the prompt learning, BAT (Cao et al., 2024) introduces a bi-directional adapter to improve multi-modal tracking based on the RGB foundation model. Considerable research has demonstrated the effectiveness and generalization of prompt learning in the RGB+X tracking community. In this work, we focus on prompt learning for the HS tracking community.

2.3. Tracking with temporal condition

Temporal condition is crucial in capturing object state changes and motion patterns (Cao et al., 2023). Consequently, mainstream studies have delved into temporal conditions in the visual tracking community (Lei et al., 2024; Wei et al., 2023; Xie et al., 2024). One common strategy involves updating the appearance representation. For instance, certain studies like Stark (Yan et al., 2021a) and SeqTrack (Chen et al., 2023b) use dynamic templates to update the object appearance for capturing changes, enhancing the match between the template and the search images. Others concentrate on learning features that encode information about the previous state or motion, such as KYS (Bhat et al., 2020) and SwinTrack (Lin et al., 2022). Additionally, integrating historical appearances is another effective approach for leveraging temporal conditions, such as UpDateNet (Zhang et al., 2019a). Recent trackers like JTBP (Lei et al., 2024), TCTrack++ (Cao et al., 2023), EVPTrack (Shi et al., 2024), TATrack (Wang et al., 2024), AQATrack (Xie et al., 2024), Seqtrack (Chen et al., 2023b), and ARTrack (Wei et al., 2023) have thoroughly demonstrated the significance of temporal condition in tracking community from various perspectives. Despite the remarkable achievements of existing RGB/RGB+X trackers, they predominantly focus on the spatial and temporal aspects, limiting their performance in complex environments where the spectral condition is crucial. Considering the substantial spectral advantages offered by HS video, we propose a unified HS video tracking framework SSTtrack that models spectral-spatial-temporal conditions in an end-to-end fashion, thereby enhancing tracking robustness in challenging scenarios.

3. Method

This section starts by overviewing our SSTtrack framework. We then delve into its key components including the multi-modal generation adapter, spectral-spatial adapter, and temporal condition adapter. Finally, we introduce the head and loss function.

3.1. Overview

As shown in Fig. 2, the proposed SSTtrack framework primarily consists of MGA, SSA, and TCA. The process begins with feeding a pair of HS patches (i.e., the search patch $X \in \mathbb{R}^{H_x \times W_x \times 16}$ and

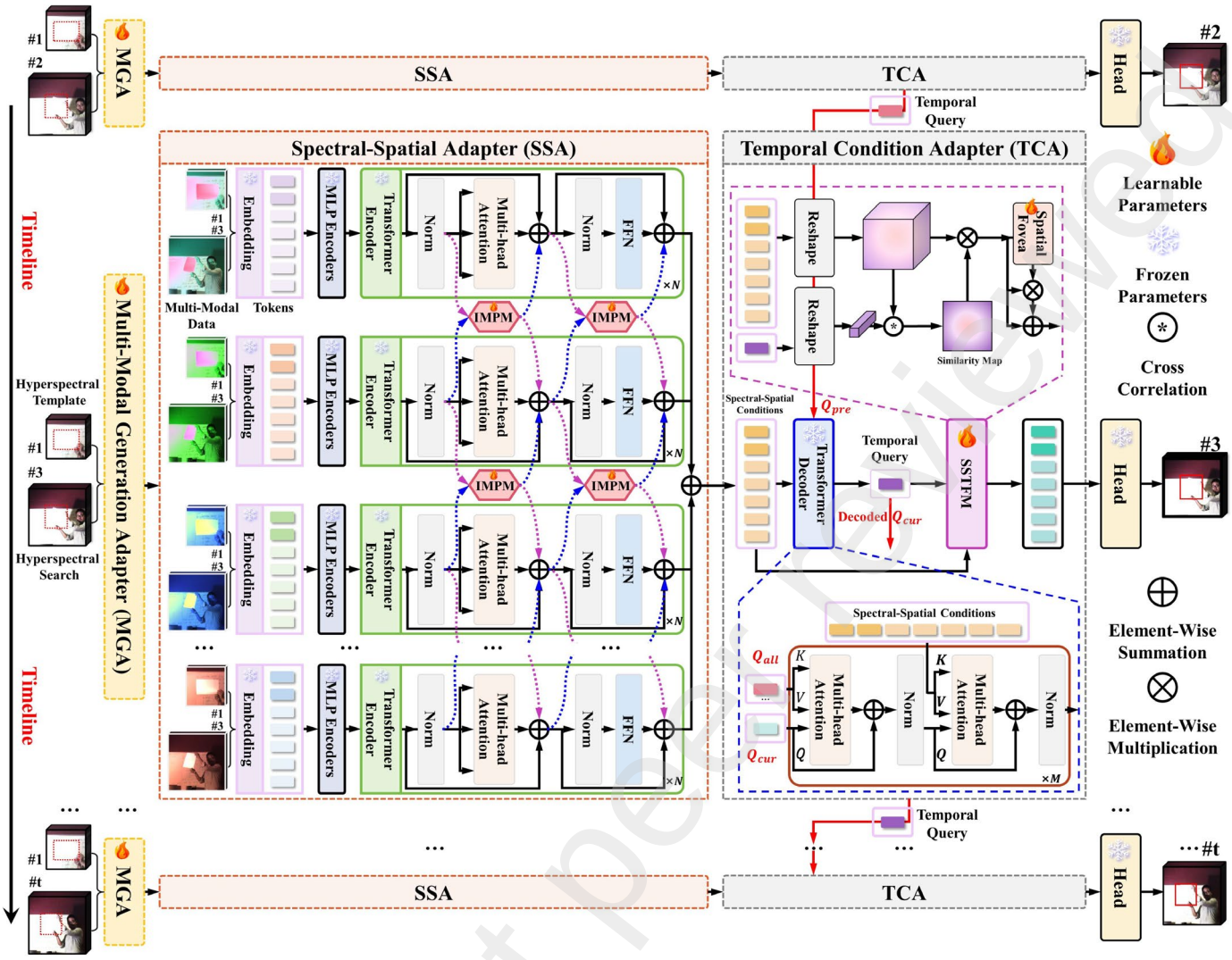

Fig. 2. Overview of the proposed SSTtrack: a unified hyperspectral video tracking framework via modeling spectral-spatial-temporal conditions. SSTtrack comprises three main components: the multi-modal generation adapter (MGA), spectral-spatial adapter (SSA), and temporal condition adapter (TCA). First, MGA achieves the multi-modal generation adaptively for bridging the band gap and stimulating the prior knowledge. Then, SSA follows a quintuple-stream architecture with an interactive modal perception module (IMPM) to learn prompts from other modalities for full exploitation of spectral and spatial conditions. While the temporal condition adapter (TCA) aims to inject the temporal condition for guiding spectral and spatial feature representations. At last, the output is fed into the head network for predicting results.

template patch $Z \in \mathbb{R}^{H_z \times W_z \times 16}$) into MGA to adaptively generate multi-modal data for capturing the rich spectral condition. Right after that, SSA dynamically extracts multi-modal complementary information using a quintuple-stream architecture, equipped with IMPM to learn feature prompts across modalities. For TCA, it has two inputs. The first one is spectral-spatial conditions derived from SSA, while the second one is several autoregressive and learnable temporal queries. It aims to inject the temporal condition to guide spectral-spatial conditions. Within TCA, the SSTFM filters spectral-spatial-temporal conditions, facilitating the integration of static and instantaneous object properties. Finally, the output features of TCA will be forwarded to the head network for result prediction.

3.2. Multi-modal generation adapter (MGA)

HS modality compiles the richer spectral condition, potentially overcoming the shortcomings of RGB modality (Li et al., 2023b). However, the scarcity of large-scale HS datasets poses a challenge for training generalized HS tracking models, especially given the data-driven models, which increases the risk of overfitting. To address this issue, we inherit the prompt learning to leverage the pre-learned knowledge embedded in RGB foundation models. Specifically, we introduce the MGA to learn the multi-modal generation by capitalizing on the interpretability benefits of integrating physical and machine models. MGA mainly consists of the inception prompt enhancing module (IPEM) and the spectral contribution prompting module (SCPM).

3.2.1. Inception prompt enhancing module (IPEM)

The self-expression model selectively identifies informative bands from C bands using the coefficient matrix $G \in \mathbb{R}^{C \times C}$ (Cai et al., 2020). The matrix reveals the relationships among bands and their contributions to downstream tracking tasks. In the HS video, each frame T is denoted as $T = [t_1, t_2, \dots, t_C] \in \mathbb{R}^{D \times C}$, where $D = m \times n$ signifies the pixel count, and $t_i \in \mathbb{R}^{m \times n}$ represents the vector for the i-th band. The HS self-expression model is thus expressed as:

$$argmin\|G\|_{1,2}, \text{s.t.}, T = TG + E, diag(G) = 0, G \geq 0$$
$$\|G\|_{1,2} = \sum_{i=1}^{C} \|g^i\|_2, \sum_{i=1}^{C} g_{ij} = 1, \forall j. \quad (1)$$

where $E \in \mathbb{R}^{D\times C}$ represents the residual item. $diag(G) = 0$ is enforced to prevent trivial solutions. $G \geq 0$ ensures that each element of g_j signifies the probability of representing t_j. $\|G\|_{1,2}$ denotes the sum of l_2-norm of all row vectors g^i. To harness the interpretability benefits of combining physical and machine models, we introduce the IPEM (Fig. 3). This module innovatively learns the solution for the matrix G of the self-expression physical model in a learning-to-optimize manner, circumventing the laborious iterative optimization process (Eq. 1). The IPEM extracts visual prompts with diverse fine-grained features, adapting to changes in object scale. Specifically, the HS image is divided into patches of size (P, P), each of which is then mapped to a 2-D tensor through a projection layer. With P values set to 18, 16, and 14, we capture features at different scales, denoted as fp_{18}, fp_{16}, and fp_{14}. Spatial max pooling Max and spatial average pooling Avg are applied to extract salient information, followed by summation, reshaping, and concatenation to obtain the token ft_{cat}. Finally, the token is fed into an MLP layer for solving solve G. This process is formulated as follows:

$$\begin{cases} ft_{18} = R\big(Max(fp_{18}) \oplus Avg(fp_{18})\big), \\ ft_{16} = R\big(Max(fp_{16}) \oplus Avg(fp_{16})\big), \\ ft_{14} = R\big(Max(fp_{14}) \oplus Avg(fp_{14})\big), \end{cases} \tag{2}$$

$$ft_{cat} = Cat(ft_{18}, ft_{16}, ft_{14}), \tag{3}$$

$$Y = MLP(MLP(ft_{cat}) \oplus ft_{cat}), \tag{4}$$

$$G = Y^T Y, \tag{5}$$

where R denotes the Reshape. Cat represents concatenation.

3.2.2. Spectral contribution prompting module (SCPM)

Upon IPEM, we derive the matrix G. Within G, the j-th column, i-th row, and (i,j)-th element are indicated by g_j, g^i, and g_{ij}, respectively. g_j denotes the coefficient of the j-th band represented by all the remaining bands, while g^i denotes the contribution of the i-th band to the reconstruction. The importance of a band is reflected by its contribution, which allows us to rank them accordingly. Thus, G unveils the relationship between bands and facilitates multi-modal generation (Chen et al., 2024a).

Specifically, G undergoes column-wise normalization to yield $\hat{g}_j = |g_{ij}|/\|g_j\|_2$, for all i, followed by $z_i = \|\hat{g}^i\|_1$, where $\hat{g}_j$ is the normalization output of the j-th column, and $\hat{g}^i$ stands for the i-th row of the normalized G. $Z \in \mathbb{R}^{C\times 1}$ indicates the desired cumulative contributions for each HS band. Subsequently, we sort all HS bands in descending order and group them to obtain multi-modal representations, denoted as $[q_1, q_2, \cdots, q_k]$, where $k = int(C/3)$ and $q_i \in \mathbb{R}^{m\times n\times 3}$.

3.3. Spectral-spatial adapter (SSA)

As depicted in Fig. 2, we propose the SSA to fully exploit spectral and spatial conditions. SSA comprises a quintuple-stream architecture with IMPM to cross-prompt multi-modal features. Each stream shares the parameters.

First, the multi-modal templates $\mathcal{L}_{tem} = [\mathcal{L}^1_{tem}, \mathcal{L}^2_{tem}, \cdots, \mathcal{L}^k_{tem}]$ and searches $L_{sea} = [\mathcal{L}^1_{sea}, \mathcal{L}^2_{sea}, \cdots, \mathcal{L}^k_{sea}]$ undergo the patch embedding and position embedding to yield tokens $x_i = [x^1_0, x^2_0, \cdots, x^k_0]$. Then, the IMPM is embedded within the i-th layer of the Transformer encoder, penetrating the encoders of multiple modalities. At each encoder's $(i+1)$ layer, it learns to fuse the modality-specific features with complementary information from other modalities, progressively refining feature prompts in a layer-by-layer fashion, as follows:

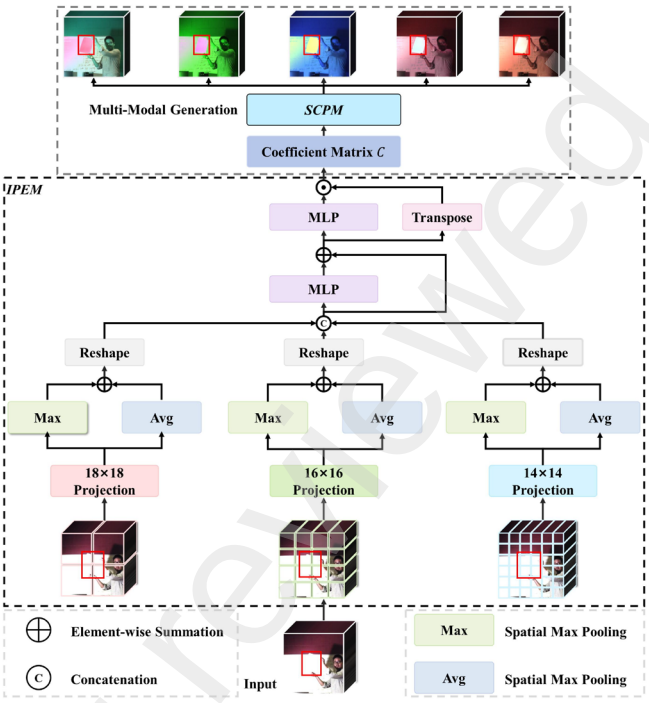

Fig. 3. Overview of the proposed multi-modal generation adapter (MGA). It consists of IPEM for solving the coefficient matrix and SCPM for generating multi-modal data.

$$\left(x^1_{i+1}, x^2_{i+1}, \cdots, x^k_{i+1}\right) = \mathcal{F}^A_i\left(x^1_i, x^2_i, \cdots, x^k_i\right), i = 1,2,\cdots,N, \tag{6}$$

where $\mathcal{F}^A_i$ means the quintuple-stream encoder layer paralleled with IMPM, The N-layer transformer encoder progressively and dynamically extracts multi-modal features of the object. Finally, the features of the multi-branch are summed.

The structure of IMPM (Fig. 4) is essentially a bottleneck design for reducing parameter burden. It comprises the down-/mid-/up-projection layers. Rectified Linear Unit (ReLU) activation serves to introduce non-linearity, while a shortcut structure is implemented to mitigate potential interference from low-level feature maps. Furthermore, a scale factor is integrated to constrain feature contributions. The output of IMPM is fed to the encoder layer of another modality as feature prompts, facilitating efficient cross-modal tracking. Taking the transfer of $x^2_i \rightarrow x^1_i$ for example, the formula is present as follows:

$$\begin{cases} x^{1'}_i = x^1_i \oplus \mathcal{F}^{Att}\left(x^1_i\right) \oplus P^2_i \\ P^2_i = \mathcal{F}^{IMP}\left(x^2_i\right), i = 1,2,\cdots,N \end{cases}, \tag{7}$$

$$\begin{cases} x^1_{i+1} = x^{1'}_i \oplus \mathcal{F}^{MLP}\left(x^{1'}_i\right) \oplus P^{2'}_i \\ P^{2'}_i = \mathcal{F}^{IMP}\left(x^{2'}_i\right), i = 1,2,\cdots,N \end{cases}, \tag{8}$$

where $\mathcal{F}^{Att}(\cdot)$, $\mathcal{F}^{IMP}(\cdot)$, and $\mathcal{F}^{MLP}(\cdot)$ mean the multi-head attention (MHA), IMPM, and multi-layer perceptron (MLP) operations, respectively. P^2_i is the feature prompt extracted derived from the x^2_i modality. The formula for $\mathcal{F}^{IMP}\left(x^2_i\right)$ is denoted as follows:

$$\mathcal{F}^{IMP}\left(x^2_i\right) = UP\left(\delta\left(MP\left(\delta\left(DP(x^2_i)\right)\right) \oplus \delta\left(DP(x^2_i)\right)\right)\right) \cdot s, \tag{9}$$

where DP, MP, and UP denote the down-projection, mid-projection, and up-projection layers, respectively. δ denotes the ReLU activation, while s is the scale factor.

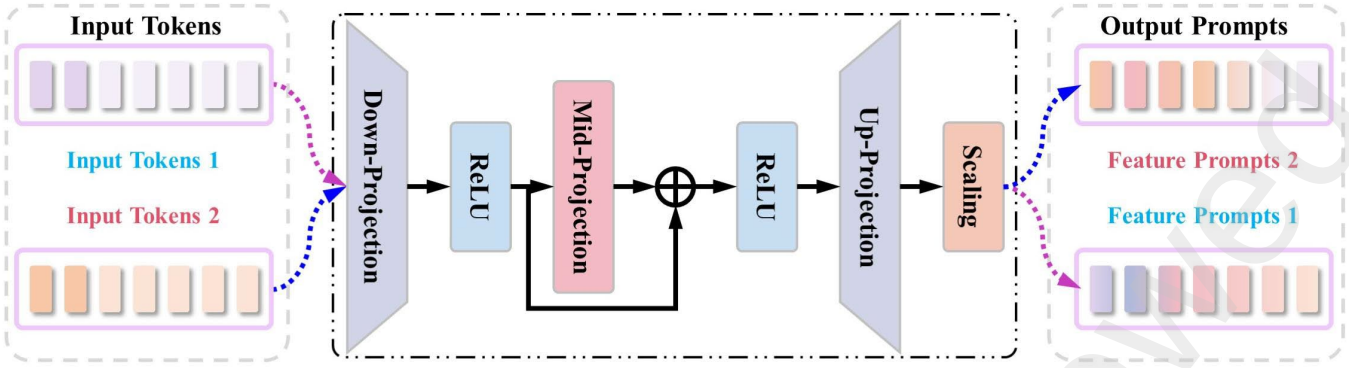

Fig. 4. Overview of the proposed interactive modal perception module (IMPM).

3.4. Temporal condition adapter(TCA)

Tracking challenges such as motion and similar objects cannot be effectively tackled when using solely spectral-spatial conditions. Therefore, we design the TCA (Fig. 2) to inject the temporal condition for guiding spectral-spatial feature representations. TCA comprises the Transformer decoder for temporal condition transfer and SSTFM for filtering spectral-spatial-temporal conditions.

The Transformer decoder has two inputs: temporal queries and spectral-spatial conditions from the SSA. There are two types of queries: Q_{pre} and Q_{cur}. Q_{pre} is passed down from previous frames. Q_{cur} is a query learned from the current frame, implying the object state while considering spectral-spatial-temporal conditions. As illustrated in Fig. 2, the Transformer decoder's queries comprise Q_{all}, Q_{cur}, and the middle parameter Q_{pre}, defined as follows:

$$Q_{all} = Cat(Q_{pre}, Q_{cur}), \tag{10}$$

$$Q_{pre} = \begin{cases} Cat(Q_1, \cdots, Q_{i-1}), t < l, \\ Cat(Q_{i-l+1}, \cdots, Q_{i-1}), t \geq l, \end{cases} \tag{11}$$

where l denotes the length of spectral-spatial-temporal conditions (tokens). The decoded Q_{cur} is propagated to the next frame of the video clip as one of the Q_{pre} in a sliding window mode. The decoder (Fig. 2) comprises M layers, each primarily consisting of two MHAs and a feed-forward neural network (FFN). The Transformer decoder interacts with historical queries to generate the decoded query, capturing temporal states and motion properties. The formula for MHA is as follows:

$$\begin{cases} MultiHead(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = Cat(\mathcal{H}_1, \mathcal{H}_2, \cdots, \mathcal{H}_{n_h})\mathcal{W}^O, \\ \mathcal{H}_i = Attention(\mathcal{Q}\mathcal{W}_i^{\mathcal{Q}}, \mathcal{K}\mathcal{W}_i^{\mathcal{K}}, \mathcal{V}\mathcal{W}_i^{\mathcal{V}}), \\ Attention(q, k, v) = Softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)v, \end{cases} \tag{12}$$

where $\mathcal{W}^O \in \mathbb{R}^{n_h \times d_m}$, $\mathcal{W}_i^{\mathcal{Q}} \in \mathbb{R}^{d_m \times d_k}$, $\mathcal{W}_i^{\mathcal{K}} \in \mathbb{R}^{d_m \times d_k}$, and $\mathcal{W}_i^{\mathcal{V}} \in \mathbb{R}^{d_m \times d_v}$ are learnable parameters. $n_h = 8$ is the number of heads. $d_m = 512$ signifies the hidden size, and $d_k = d_v = d_m/8 = 64$.

The SSTFM (Fig. 2) is designed to filter spectral, spatial, and temporal conditions with almost no learnable parameters. We first obtain the search region token f_{sea} from the encoder output. Next, we calculate the similarity between f_{sea} and the decoded token f_{tem}. The similarity map undergoes element-wisely multiplication with the search region feature followed by spatial fovea and shortcut operations to enhance the significant regions while filtering out less discriminative ones.

Subsequently, the output feature is forwarded to the head network. In summary, the process is described as follows:

$$\begin{cases} f_{sea'} = \left(R(f_{sea}) \circledast R(f_{dec})\right) \otimes f_{sea}, \\ f_{sea''} = f_{sea'} \otimes f_{fovea}, \\ f_{fovea} = \left\{ \frac{e^{f_{sea'}^{[:,i,j]}}}{\sum e^{f_{sea'}^{[:,i,j]}}} \lambda \right\}, \\ F_{sea} = f_{sea''} \oplus f_{sea'}, \end{cases} \tag{13}$$

where $i = 1,2,\cdots,H_s$ and $j = 1,2,\cdots,W_s$. H_s and W_s are the height and width of the reshaped feature of f_{sea}. $\circledast$ denotes the cross-correlation operation, and λ is a learnable weight.

3.5. Head and loss

The center-based head network is used for prediction by producing three outputs: the classification score map with $\mathbb{R}^{\frac{H_x}{P} \times \frac{W_x}{P}}$, the bounding box size with $\mathbb{R}^{2 \times \frac{H_x}{P} \times \frac{W_x}{P}}$, and the offset size with $\mathbb{R}^{2 \times \frac{H_x}{P} \times \frac{W_x}{P}}$. The object's location is determined by the highest classification score, and the final result considers both the bounding box size and offset size. We apply the weighted focal loss (Lin et al., 2017) L_{cls} for classification, L1 loss L_1 and GIoU loss (Rezatofighi et al., 2019) L_{iou} for bounding box regression, and another L1 loss L_{rec} for spectral reconstruction. The overall loss L_{tot} is obtained by:

$$L_{tot} = L_{cls} + \varphi_{L_1} L_1 + \varphi_{iou} L_{iou} + \varphi_{rec} L_{rec} \tag{14}$$

where φ_{L_1}, φ_{iou}, and φ_{rec} are regularization parameters and set to 5, 2, and 2, respectively.

4. Experimental results and analysis

4.1. Datasets

SSTtrack is trained and evaluated on HS datasets acquired from the Hyperspectral Object Tracking Competition (HOTC), with 35 sets for testing and 40 sets for training (Xiong et al., 2020). Each dataset has frame-level annotations and three types of videos: HS, false color, and RGB. The false color video is created from the corresponding HS video, while the RGB video is captured from a similar perspective. The HOTC dataset facilitates tracker performance assessment, featuring 11 attributes: Scale Variation (SV), Motion Blur (MB), Fast Motion (FM), Low Resolution (LR), Out-of-View (OV), Occlusion (OCC), Background Clutter (BC), Illumination Variation (IV), Deformation (DEF), Out-of-Plane Rotation (OPR), and In-Plane Rotation (IPR).

4.2. Implementation details

The proposed SSTtrack employs PyTorch 2.0.0 in Python 3.8 and

Table 1

Comparative results and properties of state-of-the-art RGB tracking methods. Trackers are listed in chronological order.

No.	Tracker	Venue	Feature/Backbone/Tag	RGB		FAC/HS		PreD	SucD
				Pre	Suc	Pre	Suc		
1	CSK (Henriques et al., 2012)	*ECCV 2012*	I	0.575	0.331	0.615	0.343	-4.0%	-1.2%
2	CN (Danelljan et al., 2014)	*CVPR 2014*	CN+I	0.646	0.380	0.643	0.379	0.3%	0.1%
3	SAMF (Li et al., 2015)	*ECCV 2015*	HOG+CN+I	0.693	0.418	0.660	0.388	3.3%	**3.0%**
4	DAT (Possegger et al., 2015)	*CVPR 2015*	CH	0.647	0.394	0.542	0.327	**10.5%**	**6.7%**
5	KCF (Henriques et al., 2015)	*TPAMI 2015*	HOG	0.613	0.377	0.586	0.358	2.7%	1.9%
6	SRDCF (Danelljan et al., 2015)	*ICCV 2015*	HOG	**0.846**	**0.568**	**0.830**	**0.554**	1.6%	1.4%
7	Staple (Bertinetto et al., 2016a)	*CVPR 2016*	HOG+CN	0.801	0.518	0.770	0.507	3.1%	1.1%
8	DSST (Danelljan et al., 2017)	*TPAMI 2017*	HOG+I	0.776	0.504	0.731	0.480	**4.5%**	2.4%
9	BACF (Galoogahi et al., 2017)	*ICCV 2017*	HOG	0.793	0.533	0.819	**0.544**	-2.6%	-1.1%
10	CSRDCF (Lukezic et al., 2018)	*IJCV 2018*	HOG+CN+CH	**0.870**	0.563	**0.833**	0.533	**3.7%**	**3.0%**
11	STRCF (Li et al., 2018b)	*CVPR 2018*	HOG+CN	0.829	**0.569**	**0.838**	**0.568**	-0.9%	0.1%
12	ARCF (Huang et al., 2019)	*ICCV 2019*	HOG+CN+I	0.818	**0.564**	0.798	0.542	2.0%	2.2%
13	AutoTrack (Li et al., 2020a)	*CVPR 2020*	HOG+CN+I	**0.831**	0.534	0.818	0.534	1.3%	0.0%
14	SiamRPN (Li et al., 2018a)	*CVPR 2018*	AlexNet	0.902	0.592	0.757	0.486	**14.5%**	**10.6%**
15	DaSiamRPN (Zhu et al., 2018)	*ECCV 2018*	AlexNet	0.878	0.622	0.850	0.575	2.8%	4.7%
16	ATOM (Danelljan et al., 2019)	*CVPR 2019*	ResNet-18	0.917	0.614	0.867	0.556	5.0%	5.8%
17	DiMP (Bhat et al., 2019)	*ICCV 2019*	ResNet-50	**0.944**	0.641	0.836	0.556	**10.8%**	**8.5%**
18	SiamRPN++ (Li et al., 2019)	*CVPR 2019*	ResNet-50	0.912	0.653	0.847	0.591	6.5%	6.2%
19	UpdateNet (Zhang et al., 2019a)	*ICCV 2019*	AlexNet/DaSiamRPN	0.863	0.595	0.833	0.551	3.0%	4.4%
20	SiamDW (Zhang et al., 2019b)	*CVPR 2019*	CIRNext22/SiamFC	0.872	0.565	0.812	0.529	6.0%	3.6%
21	SiamMask (Wang et al., 2019)	*CVPR 2019*	ResNet-50	0.877	0.611	0.813	0.554	6.4%	5.7%
22	PrDiMP (Danelljan et al., 2020)	*CVPR 2020*	ResNet-50	0.917	0.634	0.829	0.565	**8.8%**	6.9%
23	SiamBAN (Chen et al., 2020)	*CVPR 2020*	ResNet-50	0.853	0.610	0.863	0.587	-1.0%	2.3%
24	SiamFC++ (Xu et al., 2020)	*AAAI 2020*	GoogLeNet	0.865	0.635	0.820	0.578	4.5%	5.7%
25	KeepTrack (Mayer et al., 2021)	*ICCV 2021*	ResNet-50	**0.951**	0.656	**0.900**	0.617	5.1%	3.9%
26	SiamGAT (Guo et al., 2021)	*CVPR 2021*	GoogLeNet	0.889	0.649	0.820	0.576	6.9%	**7.3%**
27	LightTrack (Yan et al., 2021b)	*CVPR 2021*	Custom	0.814	0.593	0.761	0.530	5.3%	6.3%
28	Stark (Yan et al., 2021a)	*ICCV 2021*	ResNet-50/ST	0.900	0.637	0.814	0.579	8.6%	5.8%
29	SiamCAR (Cui et al., 2022)	*IJCV 2022*	ResNet-50	0.882	0.636	0.846	0.586	3.6%	5.0%
30	OSTrack (Ye et al., 2022)	*ECCV 2022*	ViT-Base/256	0.910	0.662	0.869	0.620	4.1%	4.2%
31	SimTrack (Chen et al., 2022)	*ECCV 2022*	ViT-Base/16	0.925	0.664	0.852	0.602	7.3%	6.2%
32	SBT (Xie et al., 2022)	*CVPR 2022*	SBT-Base	**0.938**	**0.677**	0.866	**0.626**	7.2%	5.1%
33	GRM (Gao et al., 2023c)	*CVPR 2023*	ViT-Large/320	0.921	**0.671**	**0.892**	**0.648**	2.9%	2.3%
34	SeqTrack (Chen et al., 2023b)	*CVPR 2023*	ViT-Base/256GOT	0.917	0.643	0.856	0.594	6.1%	4.9%
35	ARTrack (Wei et al., 2023)	*CVPR 2023*	ViT-Base/256	0.929	**0.673**	**0.899**	**0.640**	3.0%	3.3%
36	SMAT (Xie et al., 2024)	*WACV 2024*	MobileViTv2	0.894	0.637	0.831	0.581	6.3%	5.6%
37	AQATrack (Xie et al., 2024)	*CVPR 2024*	HiViT-Base/256	0.926	0.670	0.880	**0.640**	4.6%	3.0%
38	SSTtrack	***Ours***	HiViT-Base	*n/a*	*n/a*	**0.956**	**0.713**	*n/a*	*n/a*

Benchmark scores are color-coded for clarity: highest (**red**), second (**green**), and third (**blue**). Video categories are represented as RGB (red-green-blue), FAC (false color), and HS (hyperspectral). Pre-degradation and Suc-degradation transformations from RGB to false color are abbreviated as PreD and SucD, respectively. The dashed line separates trackers employing hand-crafted and deep features. *n/a* indicates not applicable.

is trained on an NVIDIA RTX 3090 GPU. The foundation model undergoes training for 150 epochs on large-scale RGB (three bands) data, including GOT-10k (Huang et al., 2021), TrackingNet (Muller et al., 2018), LaSOT (Fan et al., 2019), and COCO (Lin et al., 2014). The optimizer is AdamW (Loshchilov et al., 2018) with a weight decay of 10^{-4}, an initial learning rate of 4×10^{-5} for the encoder with HiViT (Zhang et al., 2022a) as the backbone, and other parameters of 4×10^{-4}. Subsequently, prompt-tuning is conducted on HS data with 16 bands for 30 epochs using the same AdamW optimizer. The initial learning rate is set to 4×10^{-4}, decayed to 4×10^{-5} at the 25-th epoch. Template and search region sizes are fixed at 128 and 256 pixels, respectively, with a batch size of 16 using video-level sampling. The tracker is tested at speed on an NVIDIA RTX 4060 GPU.

4.3. Assessment metrics

Trackers are subjected to comparative analysis through one-pass evaluations utilizing precision and success plots (Wu et al., 2015). Precision plots measure the percentage of frames where the center location error (v) falls below predefined pixel thresholds ranging from 1 to 50 pixels. v is defined as $v = \sqrt{(x-X)^2 + (y-Y)^2}$,

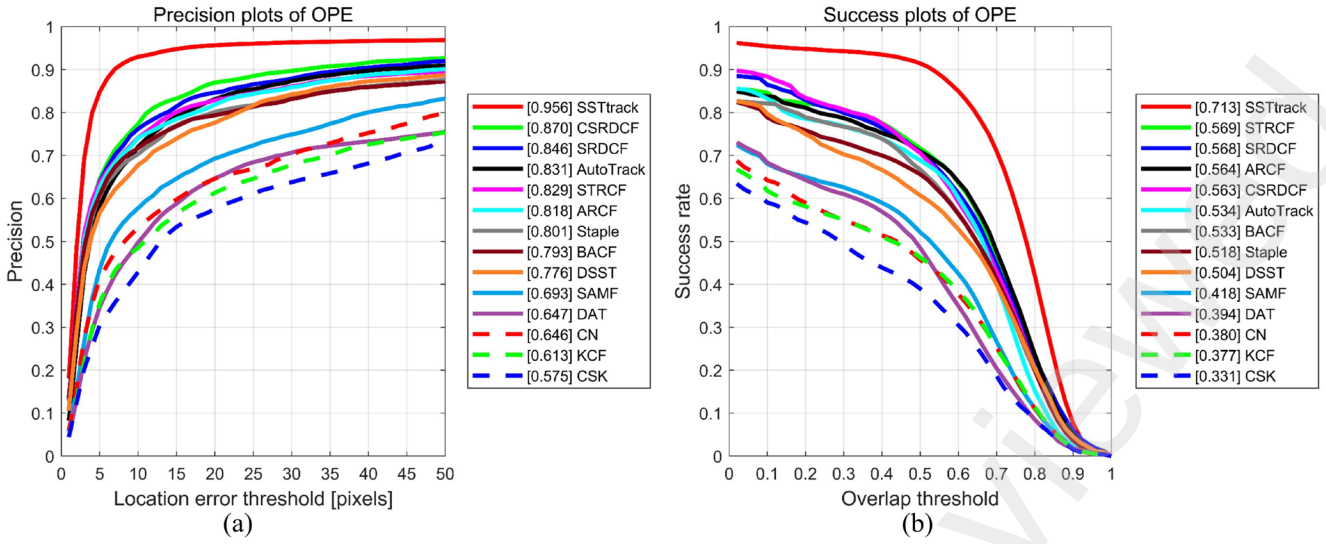

Fig. 5. Comparative results with hand-crafted feature-based trackers on RGB videos. (a) Precision plot. (b) Success plot.

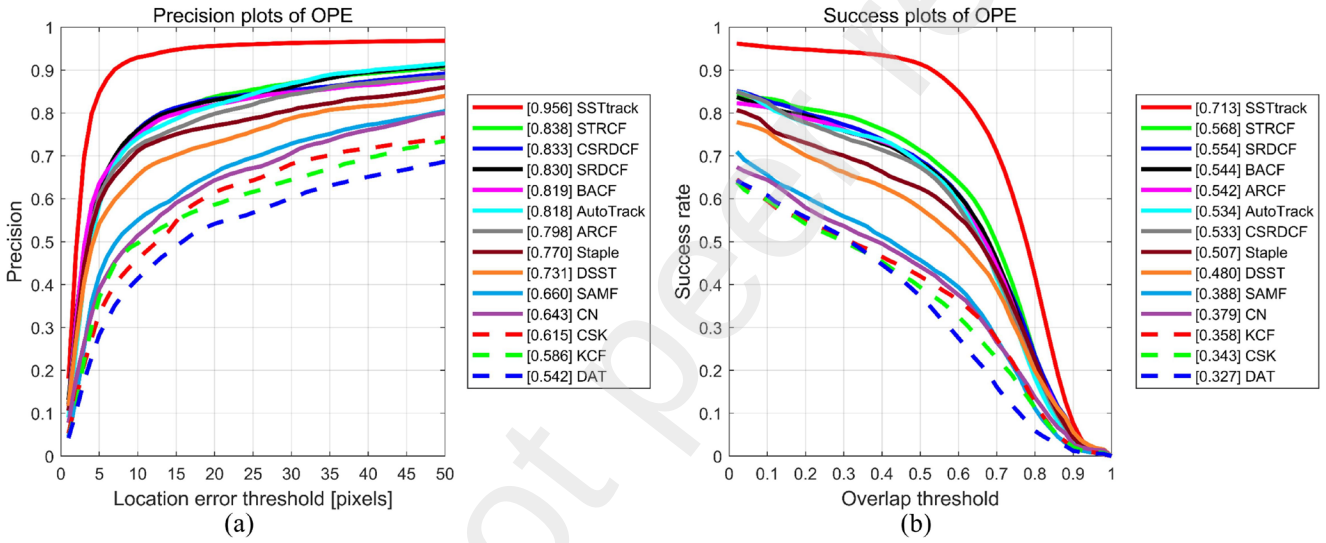

Fig. 6. Comparative results with hand-crafted feature-based trackers on false color videos. (a) Precision plot. (b) Success plot.

where (x, y) and (X, Y) are the center of the prediction r_t and the ground truth r_g. Success plots assess the percentage of successful frames where the overlap score (s) exceeds thresholds ranging from 0 to 1. s is computed using $s = |r_t \cap r_g| / |r_t \cup r_g|$, where union $\cup$ and intersection $\cap$ are mathematical set operations, and $|\cdot|$ is the pixel count in a given region. Trackers are ranked according to their precision at 20 pixels of the precision plot and the area under the curve of the success plot, i.e., Pre and Suc. The inference speed is assessed in frames per second (FPS). Moreover, the Params (M) is also present for detailed ablation comparisons.

4.4. Evaluations against leading RGB trackers

This section evaluates SSTtrack with 34 state-of-the-art (SOTA) RGB trackers, categorized into two groups: (i) trackers using hand-crafted features and (ii) those utilizing deep features.

4.4.1. Trackers employing hand-crafted features

In the benchmark analysis, we compare SSTtrack with 13 representative RGB trackers that employ hand-crafted features, namely AutoTrack (Li et al., 2020a), ARCF (Huang et al., 2019), STRCF (Li et al., 2018b), CSRDCF (Lukezic et al., 2018), BACF (Galoogahi et al., 2017), DSST (Danelljan et al., 2017), Staple (Bertinetto et al., 2016a), SRDCF (Danelljan et al., 2015), KCF (Henriques et al., 2015), DAT (Possegger et al., 2015), SAMF (Li et al., 2015), CN (Danelljan et al., 2014), and CSK (Henriques et al., 2012). Compared trackers are tested on both RGB and false-color videos, whereas SSTtrack is assessed using HS videos. Table 1 provides an overview of the characteristics and results of trackers. Precision and success plots for RGB video trials are illustrated in Fig. 5, showcasing SSTtrack's remarkable performance with a Pre of 0.956 and a Suc of 0.713. Markedly outperforming STRCF, SSTtrack achieves a 12.7% enhancement in Pre and a 14.4% lift in Suc. Furthermore, it outperforms SRDCF and ARCF by 14.5% and 14.9% in Suc, respectively, underscoring the potential of harnessing the rich spectral-spatial-temporal conditions inherent in HS data. Transforming the HS video into a false color representation facilitates the feasible implementation of the RGB tracker. Subsequently, we proceed to perform experiments utilizing the false color video, as depicted in Fig. 6 and Table 1. It is observed that STRCF demonstrates notable results, with SRDCF, BACF, and ARCF slightly trailing behind. However, SSTtrack stands out as the frontrunner, surpassing them by margins of 14.5%, 15.9%, 16.9%, and 17.1%, respectively.

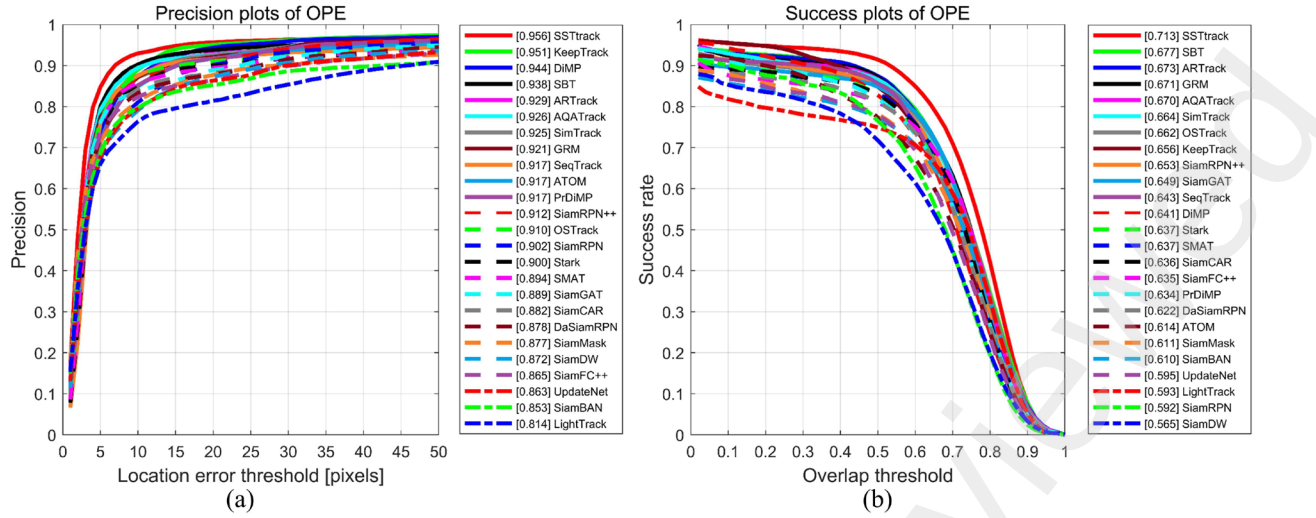

Fig. 7. Comparative results with deep feature-based trackers on RGB videos. (a) Precision plot. (b) Success plot.

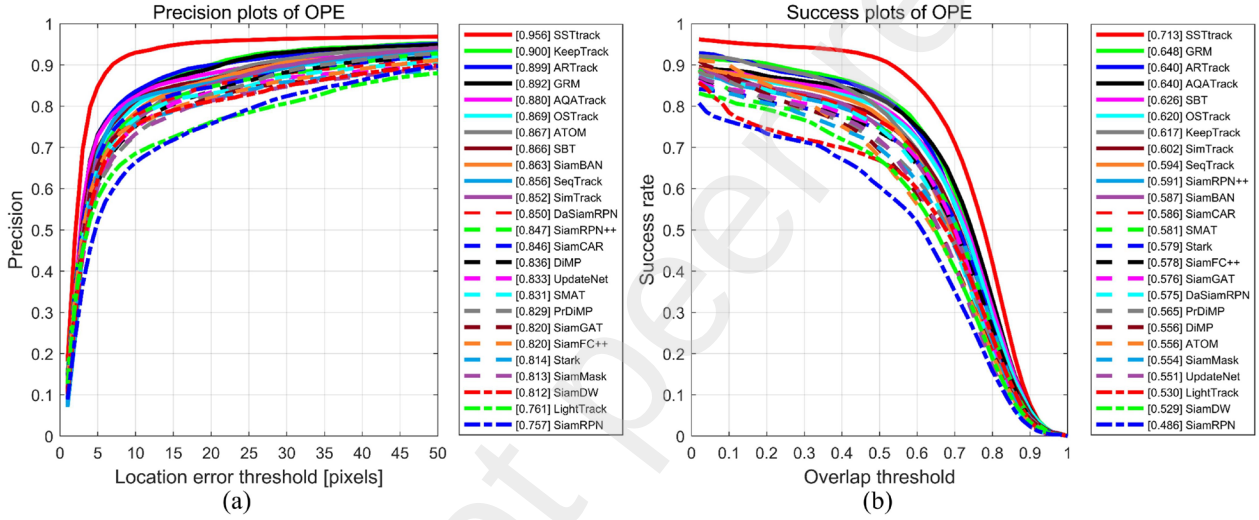

Fig. 8. Comparative results with deep feature-based trackers on false color videos. (a) Precision plot. (b) Success plot.

4.4.2. Trackers employing deep features

In the realm of RGB tracking, deep features, renowned for their discriminative capabilities, have garnered considerable attention (Javed et al., 2022). Here, SSTtrack undergoes evaluation against 24 SOTA deep feature-based trackers, involving AQATrack (Xie et al., 2024), SMAT (Yelluru Gopal et al., 2024), ARTrack (Wei et al., 2023), SeqTrack (Chen et al., 2023b), GRM (Gao et al., 2023c), SBT (Xie et al., 2022), SimTrack (Chen et al., 2022), OSTrack (Ye et al., 2022), SiamCAR (Cui et al., 2022), Stark (Yan et al., 2021a), LightTrack (Yan et al., 2021b), SiamGAT (Guo et al., 2021), KeepTrack (Mayer et al., 2021), SiamFC++ (Xu et al., 2020), SiamBAN (Chen et al., 2020), PrDiMP (Danelljan et al., 2020), SiamMask (Wang et al., 2019), SiamDW (Zhang et al., 2019b), UpdateNet (Zhang et al., 2019a), SiamRPN++ (Li et al., 2019), DiMP (Bhat et al., 2019), ATOM (Danelljan et al., 2019), DaSiamRPN (Zhu et al., 2018), and SiamRPN (Li et al., 2018a). Table 1 showcases a detailed overview of the results and properties. Figs. 7 and 8 depict the precision plot and success plot for experiments conducted on RGB and false color videos, respectively. Our tracker consistently maintains the leading performance across 24 SOTA trackers. Compared to the baseline AQATrack tested on false color videos, our SSTtrack exhibits gains of 7.6% in Pre and 7.3% in Suc, respectively. Especially, SSTtrack outperforms top-ranked KeepTrack and SBT by 5.7% and 3.6%, respectively, in terms of Suc on RGB videos, and by 9.6% and 8.7% on false color videos. To summarize, the exhaustive results validate the impressive performance of our SSTtrack.

4.5. Parallel analysis with RGB trackers

Fig. 9 depicts the parallel analysis of the top 13 trackers utilizing both hand-crafted and deep features across RGB and false color videos. Through the analysis of Fig. 9, we have gleaned some valuable insights that will enrich our comprehension in the realm of HS video tracking. First, it is noteworthy that trackers employing deep features often outperform those relying on hand-crafted features, regardless of RGB or false color video. This superiority arises from the capability of deep models to acquire generalized representations from training sets. Additionally, the performance of RGB trackers generally diminishes on false color videos compared to RGB videos, albeit with relatively consistent rankings. Third, trackers utilizing deep features demonstrate more pronounced performance degradation, measured in terms of PreD and SucD, compared to their counterparts relying on hand-crafted features. This decline can be attributed to the heavy dependence on properties present in RGB format training datasets. Remarkably, our tracker achieves superior performance by effectively leveraging the spectral-

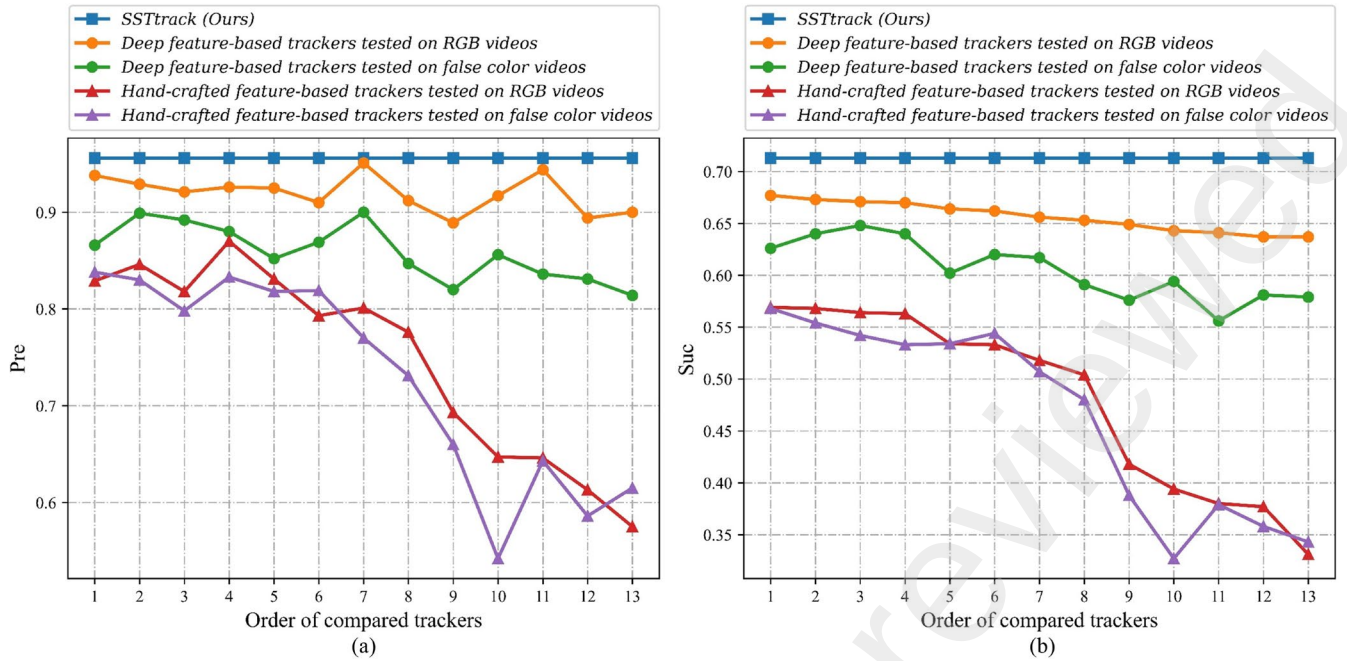

Fig. 9. Parallel analysis alongside RGB trackers. (a) Pre score. (b) Suc score. The order of compared trackers is determined by their Suc scores on the RGB video. Trackers employing hand-crafted features are listed as follows: STRCF, SRDCF, ARCF, CSRDCF, AutoTrack, BACF, Staple, DSST, SAMF, DAT, CN, KCF, and CSK. Trackers employing deep features are listed as follows: SBT, ARTrack, GRM, AQATrack, SimTrack, OSTrack, KeepTrack, SiamRPN++, SiamGAT, SeqTrack, DiMP, SMAT, and Stark.

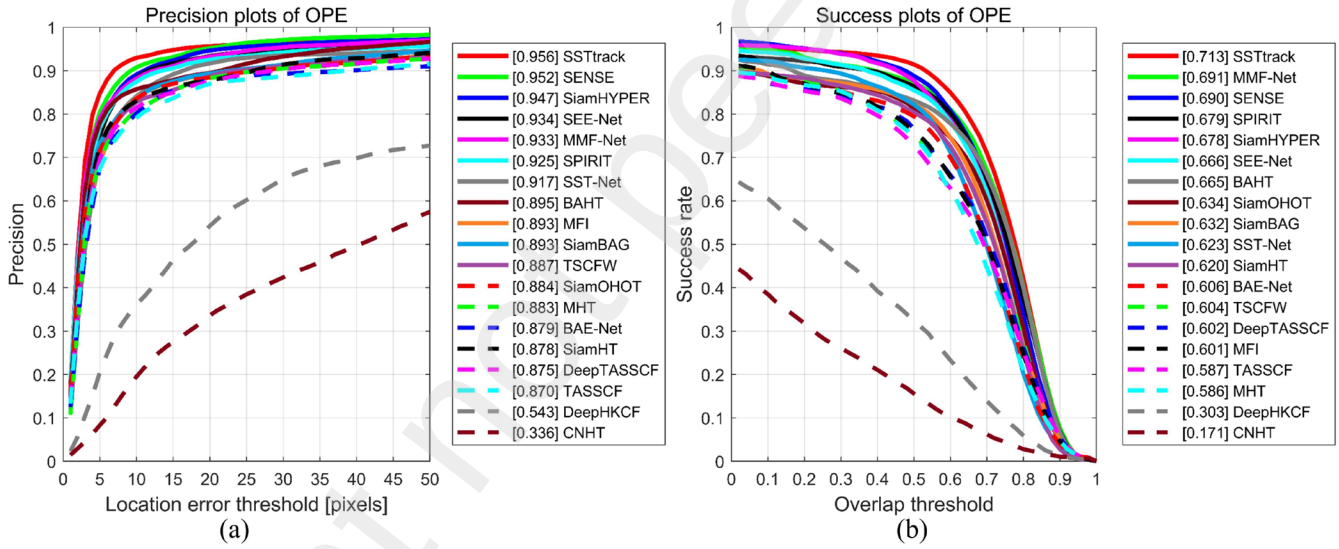

Fig. 10. Comparative results with HS trackers on HS videos. (a) Precision plot. (b) Success plot.

spatial-temporal conditions inherent in HS videos. In addition, the transformation of HS video to a false color format unavoidably leads to the loss and distortion of critical spectral-spatial conditions, thereby impacting the stability of trackers.

4.6. Comparison with HS trackers

Here, we analyze SSTtrack in comparison to 18 representative HS trackers, i.e., MMF-Net (Li et al., 2024), SENSE (Chen et al., 2024a), SPIRIT (Chen et al., 2024b), SiamHT (Tang et al., 2023), SiamBAG (Li et al., 2023a), SiamOHOT (Sun et al., 2023), SEE-Net (Li et al., 2023b), DeepTASSCF (Tang et al., 2022a), TASSCF (Tang et al., 2022a), BAHT (Tang et al., 2022b), TSCFW (Hou et al., 2022), SiamHYPER (Liu et al., 2022), SST-Net (Li et al., 2021), MFI (Zhang et al., 2021), BAE-Net (Li et al., 2020b), MHT (Xiong et al., 2020), DeepHKCF (Uzkent et al., 2019), and CNHT (Qian et al., 2018). Table 2 shows the results and properties of these trackers, while Fig. 10 depicts the precision plot and success plot. Notably, in the precision plot analysis, SENSE, SiamHYPER, SEE-Net, MMF-Net, SPIRIT, and SST-Net stand out as the top performers, achieving Pre scores of 0.952, 0.947, 0.934, 0.933, 0.925, and 0.917, respectively. In contrast, SSTtrack showcases even superior results, achieving a score of 0.956, surpassing them by margins of 0.4%, 0.9%, 2.2%, 2.3%, 3.1%, and 3.9%, respectively. Regarding the success plot analysis, MMF-Net, SENSE, SPIRIT, SiamHYPER, SEE-Net, and BAHT show promising performance, with Suc scores of 0.691, 0.690, 0.679, 0.678, 0.666, and 0.665, respectively. Our tracker obtains notable improvements over them by up to 2.2%, 2.3%, 3.4%, 3.5%, 4.7%, and 4.8%, respectively. In summary, SSTtrack demonstrates superior performance in both Pre and Suc scores, highlighting the effectiveness of modeling spectral-spatial-temporal conditions in the HS tracking domain.

Table 2

Results and properties of comparative HS trackers tested on HS videos. Trackers are listed in chronological order.

No.	Tracker	Venue	Framework	Feature	UFB	UTC	MOP	FPS	Pre	Suc
1	CNHT (Qian et al., 2018)	*ICSM 2018*	KCF	Deep feature	✓	-	CPU	2.6	0.336	0.171
2	DeepHKCF (Uzkent et al., 2019)	*TGRS 2019*	KCF	Deep feature	-	-	CPU	0.9	0.543	0.303
3	MHT (Xiong et al., 2020)	*TIP 2019*	KCF	Hand-crafted feature	✓	-	CPU	2.2	0.883	0.586
4	BAE-Net (Li et al., 2020b)	*ICIP 2020*	VITAL	Deep feature	✓	-	GPU	0.5	0.879	0.606
5	MFI (Zhang et al., 2021)	*WISP 2021*	KCF	Deep + Hand-crafted features	-	-	CPU	0.4	0.893	0.601
6	SST-Net (Li et al., 2021)	*WISP 2021*	VITAL	Deep feature	✓	✓	GPU	0.5	0.917	0.623
7	SiamHYPER (Liu et al., 2022)	*TIP 2022*	SiamFC	Deep feature	✓	-	GPU	**19.0**	**0.947**	0.678
8	TSCFW (Hou et al., 2022)	*TGRS 2022*	KCF	Hand-crafted features	✓	-	CPU	3.4	0.887	0.604
9	BAHT (Tang et al., 2022b)	*GRSL 2022*	SiamFC	Deep feature	-	-	GPU	16.0	0.895	0.665
10	TASSCF (Tang et al., 2022a)	*CVIU 2022*	KCF	Hand-crafted features	-	-	CPU	16.0	0.870	0.587
11	DeepTASSCF (Tang et al., 2022a)	*CVIU 2022*	KCF	Deep feature	-	-	CPU	6.0	0.875	0.602
12	SEE-Net (Li et al., 2023b)	*TIP 2023*	SiamFC	Deep feature	✓	-	GPU	8.7	0.934	0.666
13	SiamOHOT (Sun et al., 2023)	*TGRS 2023*	SiamFC	Deep feature	✓	-	GPU	**38.0**	0.884	0.634
14	SiamBAG (Li et al., 2023a)	*TGRS 2023*	SiamFC	Deep feature	✓	-	GPU	5.7	0.893	0.632
15	SiamHT (Tang et al., 2023)	*NCA 2023*	SiamFC	Deep feature	-	-	GPU	16.0	0.878	0.620
16	SPIRIT (Chen et al., 2024b)	*TGRS 2024*	SiamFC	Deep feature	✓	✓	GPU	**26.0**	0.925	0.679
17	SENSE (Chen et al., 2024a)	*IF 2024*	SiamFC	Deep feature	✓	✓	GPU	15.4	**0.952**	**0.690**
18	MMF-Net (Li et al., 2024)	*TGRS 2024*	SiamFC	Deep feature	✓	-	GPU	6.1	0.933	**0.691**
19	SSTtrack	***Ours***	SiamFC	Deep feature	✓	✓	GPU	15.5	**0.956**	**0.713**

For the Venue, WISP corresponds to the WHISPERS, and IF refers to the Information Fusion. Regarding the Framework, KCF (Henriques et al., 2015) signifies the kernelized correlation filter. VITAL (Song et al., 2018) involves visual tracking via adversarial learning. SiamFC (Bertinetto et al., 2016b) is a fully convolutional Siamese network. UFB signifies an effort to utilize the full band, while UTC denotes an exploration into leveraging temporal conditions. MOP stands for the main operation platform.

4.7. Accuracy vs. speed on HS videos

Fig. 11 depicts the trade-off between accuracy (Pre and Suc) and speed (FPS) for various HS trackers. The competitors encompass MHT (Xiong et al., 2020), BAE-Net (Li et al., 2020b), MFI (Zhang et al., 2021), SST-Net (Li et al., 2021), SiamHYPER (Liu et al., 2022), TSCFW (Hou et al., 2022), BAHT (Tang et al., 2022b), TASSCF (Tang et al., 2022a), DeepTASSCF (Tang et al., 2022a), SEE-Net (Li et al., 2023b), SiamOHOT (Sun et al., 2023), SiamBAG (Li et al., 2023a), SiamHT (Tang et al., 2023), SPIRIT (Chen et al., 2024b), SENSE (Chen et al., 2024a), MMF-Net (Li et al., 2024), and our SSTtrack. In Fig. 11(a), it is evident that most HS trackers either exhibit speeds lower than the average FPS value of 11.496 (e.g., SEE-Net, MMF-Net, and SST-Net) or accuracy lower than the average Pre value of 0.889 (e.g., BAHT, SiamOHOT, and SiamHT). Only SSTtrack, SENSE, SiamHYPER, and SPIRIT surpass both theaverage Pre and FPS levels. Furthermore, our SSTtrack yields the highest Pre score of 0.956. In Fig. 11(b), SSTtrack, SENSE, SPIRIT, SiamHYPER, and BAHT exceed the average Suc of 0.640 and FPS. SSTtrack notably achieves the highest Suc of 0.713, surpassing SENSE and SPIRIT by 2.3% and 3.4%, respectively. In summary, SSTtrack strikes a favorable trade-off between accuracy and speed, making it a solid candidate for HS video tracking.

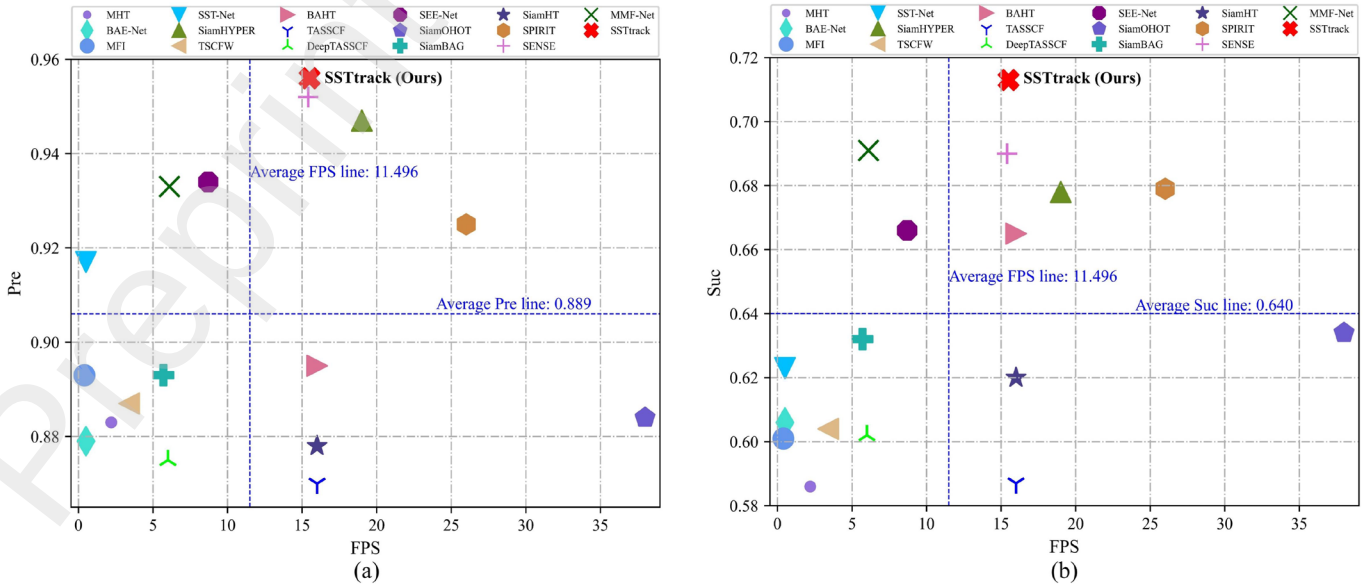

Fig. 11. Comparison of accuracy and speed with HS trackers on HS videos. (a) Pre score vs. FPS value. (b) Suc score vs. FPS value.

Table 3

Pre scores for each attribute and overall performance.

No.	Tracker	Venue	BC	DEF	FM	IPR	IV	LR	MB	OCC	OPR	OV	SV	OVE
1	STRCF	*CVPR 2018*	0.801	**0.964**	0.833	0.924	0.848	0.705	0.853	0.812	0.926	0.887	0.834	0.838
2	ARCF	*ICCV 2019*	0.740	0.910	0.792	0.899	0.771	0.671	0.799	0.740	0.878	0.887	0.806	0.798
3	AutoTrack	*CVPR 2020*	0.759	0.930	0.737	0.882	0.810	0.700	0.827	0.788	0.888	0.891	0.825	0.818
4	LightTrack	*CVPR 2021*	0.673	0.924	0.848	0.793	0.687	0.688	0.928	0.722	0.850	0.896	0.768	0.761
5	SiamCAR	*IJCV 2022*	0.800	0.935	0.952	0.860	0.868	0.855	0.931	0.818	0.905	0.887	0.877	0.846
6	SeqTrack	*CVPR2023*	0.831	0.935	0.968	0.896	0.809	0.827	0.929	0.860	0.960	**0.995**	0.883	0.856
7	SMAT	*WACV 2024*	0.761	0.951	0.957	0.871	0.828	0.778	0.959	0.823	0.944	0.891	0.868	0.831
8	AQATrack	*CVPR2024*	0.869	0.950	**1.000**	0.944	0.844	0.796	0.984	0.870	0.959	**1.000**	0.889	0.880
9	MHT	*TIP 2019*	0.901	0.908	0.774	0.940	0.806	0.827	0.839	0.816	0.893	0.887	0.874	0.883
10	BAE-Net	*ICIP 2020*	0.921	0.940	0.871	**0.985**	0.824	0.740	0.882	0.794	**0.982**	0.896	0.892	0.879
11	MFI	*WISP 2021*	0.929	0.885	0.832	0.952	**0.887**	0.878	0.841	0.815	0.950	**0.928**	0.916	0.893
12	SST-Net	*WISP 2021*	**0.980**	**0.971**	0.835	**0.981**	0.817	0.746	0.832	0.860	**0.980**	0.896	0.907	0.917
13	TSCFW	*TGRS 2022*	0.907	0.901	0.829	0.956	0.859	0.885	0.853	0.810	0.929	0.896	0.902	0.887
14	BAHT	*GRSL 2022*	0.930	0.942	0.989	0.970	0.760	0.823	**0.988**	0.842	0.964	0.891	0.881	0.895
15	TASSCF	*CVIU 2022*	0.906	0.914	0.792	0.919	0.873	0.851	0.818	0.781	0.921	0.882	0.870	0.870
16	DeepTASSCF	*CVIU 2022*	0.903	0.911	0.794	0.948	0.851	0.839	0.860	0.794	0.949	0.891	0.893	0.875
17	SEE-Net	*TIP 2023*	0.959	0.938	0.991	**0.983**	0.879	**0.943**	0.981	**0.888**	0.976	0.891	**0.929**	**0.934**
18	SiamOHOT	*TGRS 2023*	0.931	0.935	0.816	0.960	0.825	0.781	0.819	0.800	0.959	0.887	0.895	0.884
19	SiamBAG	*TGRS 2023*	0.899	0.936	0.883	0.930	0.846	0.839	0.899	0.831	0.906	0.891	0.892	0.893
20	SiamHT	*NCA 2023*	0.878	0.927	0.890	0.951	0.875	0.868	0.881	0.829	0.954	0.896	0.911	0.878
21	SENSE	*IF 2024*	**0.966**	0.948	**0.992**	0.976	**0.922**	**0.946**	**0.993**	**0.921**	0.971	0.891	**0.946**	**0.952**
22	SSTtrack	***Ours***	**0.968**	**0.990**	**0.995**	**0.981**	**0.932**	**0.948**	**0.989**	**0.929**	**0.989**	**0.995**	**0.948**	**0.956**

Table 4

Suc scores for each attribute and overall performance.

No.	Tracker	Venue	BC	DEF	FM	IPR	IV	LR	MB	OCC	OPR	OV	SV	OVE
1	STRCF	*CVPR 2018*	0.546	0.681	0.577	0.670	0.531	0.411	0.600	0.564	0.675	0.624	0.568	0.568
2	ARCF	*ICCV 2019*	0.513	0.648	0.577	0.643	0.485	0.426	0.552	0.510	0.637	0.616	0.537	0.542
3	AutoTrack	*CVPR 2020*	0.512	0.659	0.553	0.618	0.473	0.421	0.546	0.513	0.631	0.616	0.528	0.534
4	LightTrack	*CVPR 2021*	0.472	0.686	0.606	0.600	0.418	0.441	0.680	0.517	0.638	0.623	0.522	0.530
5	SiamCAR	*IJCV 2022*	0.565	0.697	0.662	0.650	0.543	0.553	0.662	0.570	0.678	0.634	0.595	0.586
6	SeqTrack	*CVPR2023*	0.566	0.693	0.638	0.664	0.553	0.528	0.610	0.595	0.710	**0.756**	0.621	0.594
7	SMAT	*WACV 2024*	0.548	0.729	0.632	0.663	0.515	0.505	0.688	0.578	0.713	0.652	0.597	0.581
8	AQATrack	*CVPR2024*	0.633	**0.739**	0.691	0.732	**0.587**	0.555	0.704	**0.635**	**0.747**	**0.828**	0.648	0.640
9	MHT	*TIP 2019*	0.606	0.664	0.542	0.670	0.477	0.475	0.560	0.564	0.644	0.626	0.574	0.586
10	BAE-Net	*ICIP 2020*	0.651	0.679	0.607	0.699	0.524	0.489	0.593	0.554	0.701	0.516	0.604	0.606
11	MFI	*WISP 2021*	0.651	0.639	0.600	0.692	0.516	0.514	0.570	0.546	0.680	0.611	0.599	0.601
12	SST-Net	*WISP 2021*	0.685	0.699	0.561	0.696	0.502	0.462	0.535	0.594	0.698	0.480	0.602	0.623
13	TSCFW	*TGRS 2022*	0.636	0.648	0.591	0.724	0.535	0.548	0.561	0.556	0.685	0.654	0.603	0.604
14	BAHT	*GRSL 2022*	**0.707**	**0.741**	0.704	**0.762**	0.514	0.554	**0.735**	0.603	**0.760**	0.582	**0.652**	0.665
15	TASSCF	*CVIU 2022*	0.606	0.646	0.575	0.675	0.541	0.485	0.589	0.538	0.666	0.584	0.573	0.587
16	DeepTASSCF	*CVIU 2022*	0.630	0.666	0.567	0.720	0.520	0.488	0.595	0.551	0.715	0.616	0.610	0.602
17	SEE-Net	*TIP 2023*	0.705	0.710	**0.711**	0.742	0.566	**0.626**	0.702	0.622	0.729	0.633	0.649	**0.666**
18	SiamOHOT	*TGRS 2023*	0.699	0.715	0.560	0.732	0.517	0.497	0.610	0.556	0.728	0.582	0.627	0.634
19	SiamBAG	*TGRS 2023*	0.648	0.691	0.614	0.703	0.533	0.582	0.649	0.597	0.683	0.634	0.622	0.632
20	SiamHT	*NCA 2023*	0.629	0.682	0.627	0.702	0.581	0.576	0.629	0.580	0.702	0.650	0.635	0.620
21	SENSE	*IF 2024*	**0.721**	0.708	**0.746**	**0.752**	**0.609**	**0.622**	**0.730**	**0.645**	0.737	0.629	**0.674**	**0.690**
22	SSTtrack	***Ours***	**0.734**	**0.782**	**0.735**	**0.774**	**0.655**	**0.665**	**0.744**	**0.673**	**0.795**	**0.770**	**0.709**	**0.713**

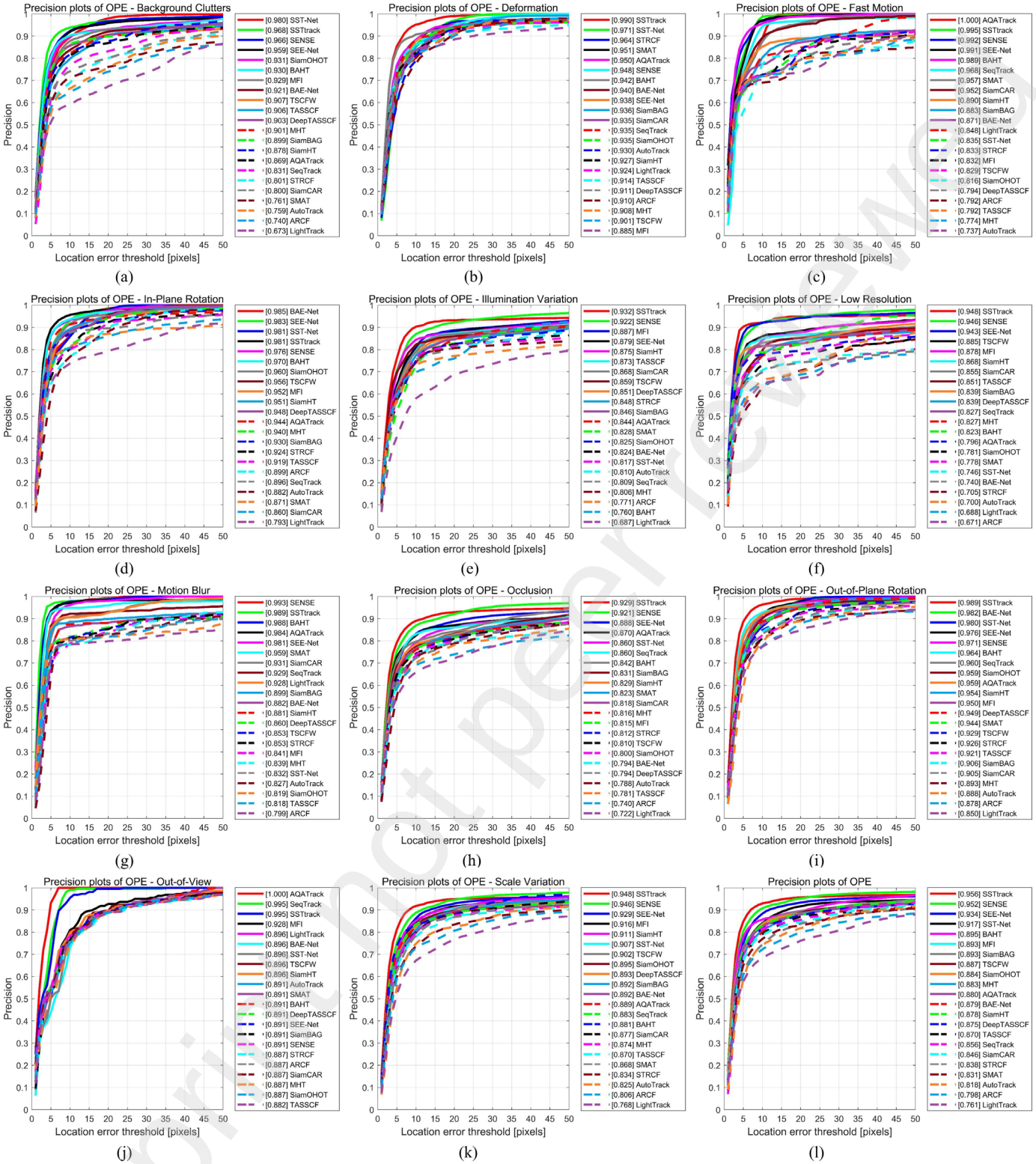

Fig. 12. Precision plots for each attribute and overall performance. (a) BC. (b) DEF. (c) FM. (d) IPR. (e) IV. (f) LR. (g) MB. (h) OCC. (i) OPR. (j) OV. (k) SV. (l) OVE.

4.8. Attribute-based evaluation

We further conduct attribute-based comparisons involving eight top-performing RGB trackers: STRCF (Li et al., 2018b), ARCF (Huang et al., 2019), AutoTrack (Li et al., 2020a), LightTrack (Yan et al., 2021b), SiamCAR (Cui et al., 2022), SeqTrack (Chen et al., 2023b), SMAT (Yelluru Gopal et al., 2024), and AQATrack (Xie et al., 2024), alongside 13 leading HS trackers: MHT (Xiong et al., 2020), BAE-Net (Li et al., 2020b), MFI (Zhang et al., 2021), SST-Net (Li et al., 2021), TSCFW (Hou et al., 2022), BAHT (Tang et al., 2022b), TASSCF (Tang et al., 2022a), DeepTASSCF (Tang et al., 2022a), SEE-Net (Li et al., 2023b), SiamOHOT (Sun et al., 2023), SiamBAG (Li et al., 2023a), SiamHT (Tang et al., 2023), and SENSE (Chen et al., 2024a). While the RGB tracker's evaluation is based on false color videos generated from HS videos, others are directly assessed on HS videos. Detailed breakdowns of Pre and Suc scores for each attribute and overall (OVE) are provided in Table 3 and

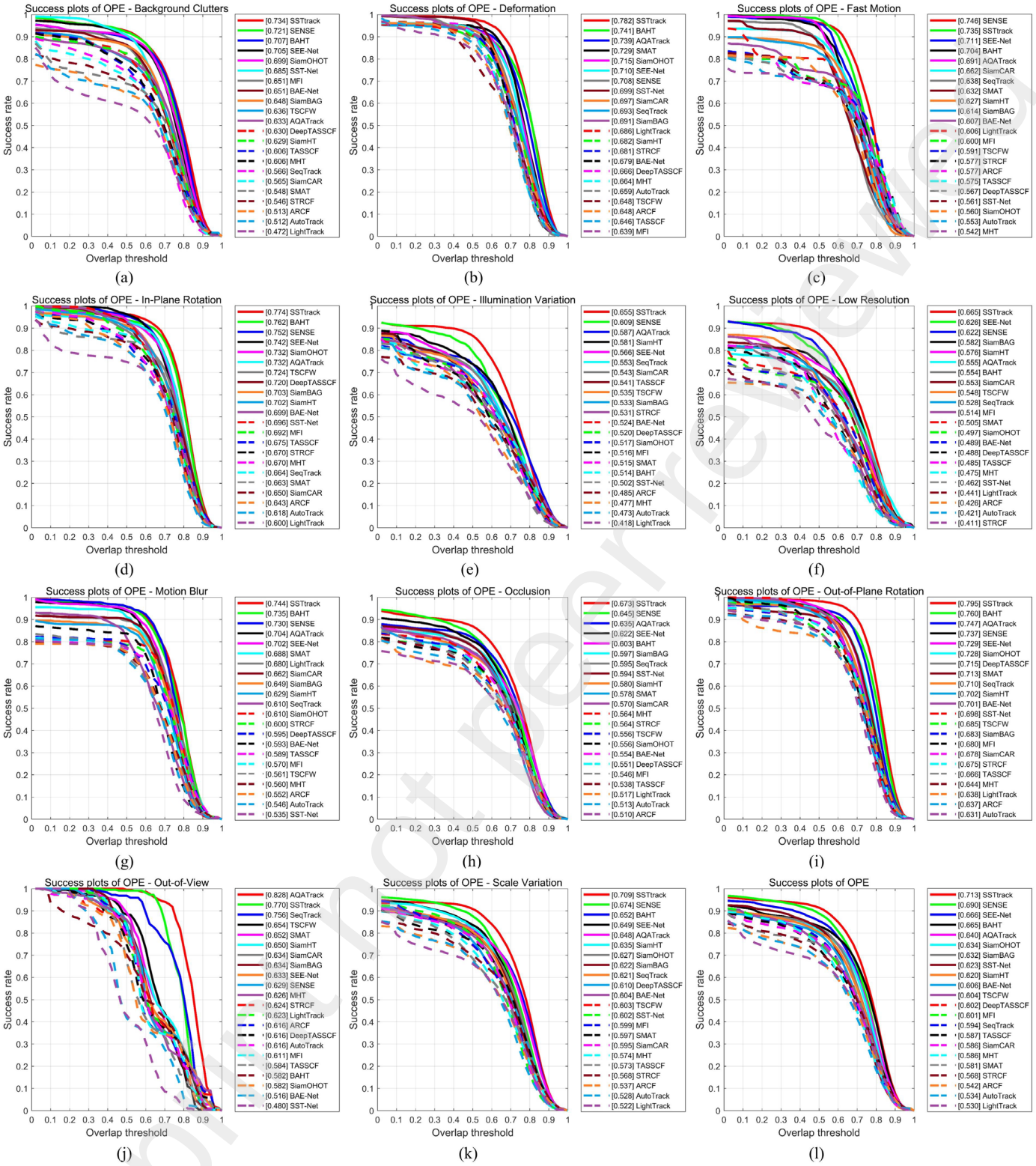

Fig. 13. Success plots for each attribute and overall performance. (a) BC. (b) DEF. (c) FM. (d) IPR. (e) IV. (f) LR. (g) MB. (h) OCC. (i) OPR. (j) OV. (k) SV. (l) OVE.

Table 4. Additionally, Fig. 12 illustrates the precision plots, while Fig. 13 showcases the success plots. Notably, in terms of the Pre score, as depicted in the detailed results, SSTtrack consistently ranks in the top three across all 11 attributes and secures the first position overall. In particular, SSTtrack holds the first place in six challenging attributes: DEF, IV, LR, OCC, OPR, and SV. Regarding the Suc score, SSTtrack ranks in the first position in nearly all attributes, except for FM and OV. Leveraging efficient and effective components, our tracker exhibits impressive performance in handling challenging attributes, thereby offering a robust benchmark for the HS tracking community.

4.9. Visual comparison

Fig. 14 displays six visual samples: *basketball*, *coke*, *excavator*, *rider1*, *student*, and *trucker*, tracked by various methods including SST-Net (Li et al., 2021), SENSE (Chen et al., 2024a), SMAT

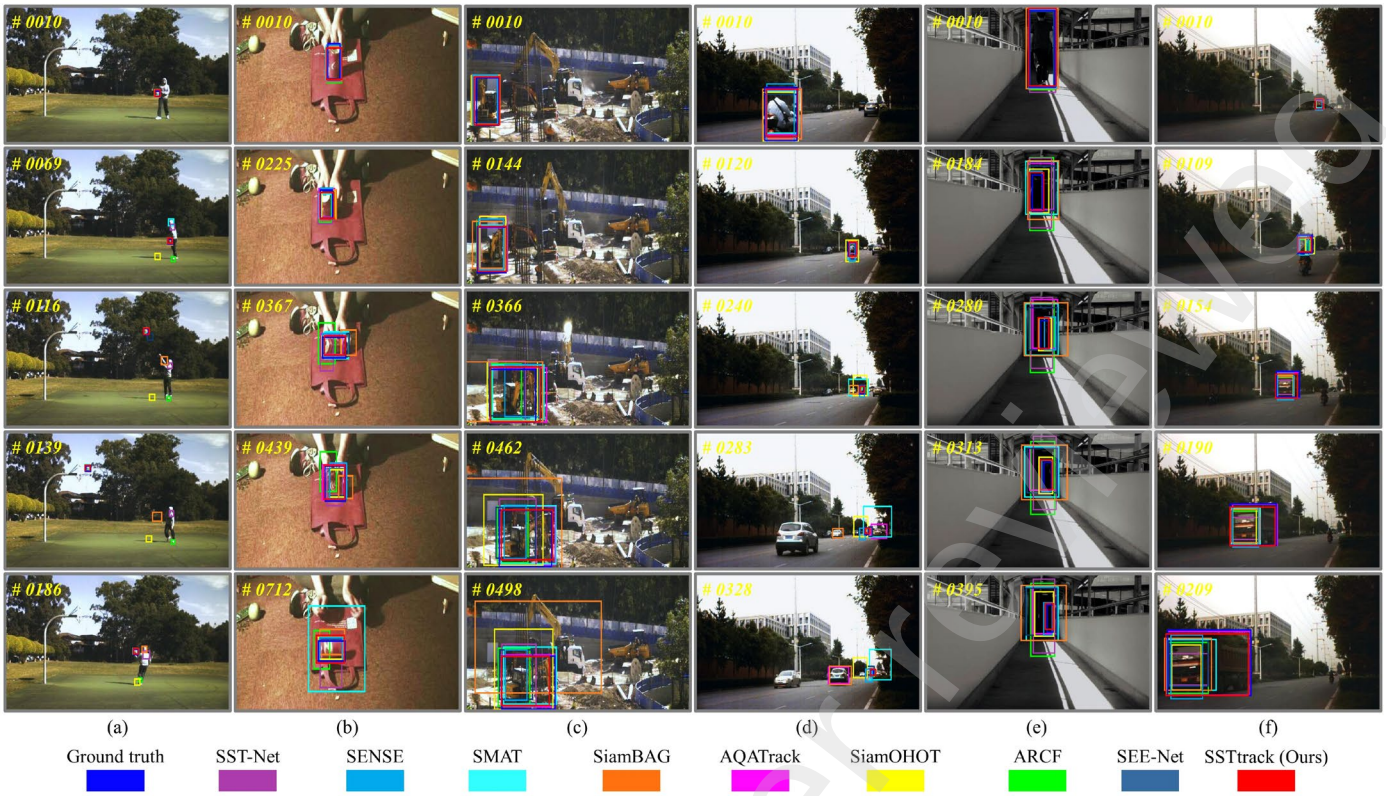

Fig. 14. Qualitative results. (a) *basketball*, attribute: MB, FM, LR, and OCC. (b) *coke*, attribute: IPR, BC, OPR, FM, and SV. (c) *excavator*, attribute: IPR, SV, OPR, OCC, and DEF. (d) *rider1*, attribute: LR, OCC, IV, and SV. (e) *student*, attribute: SV and IV. (f) *trucker*, attribute: OV, OCC, IV, and SV. Bounding boxes are visualized in the false color image, with the current frame displayed in the top-left corner.

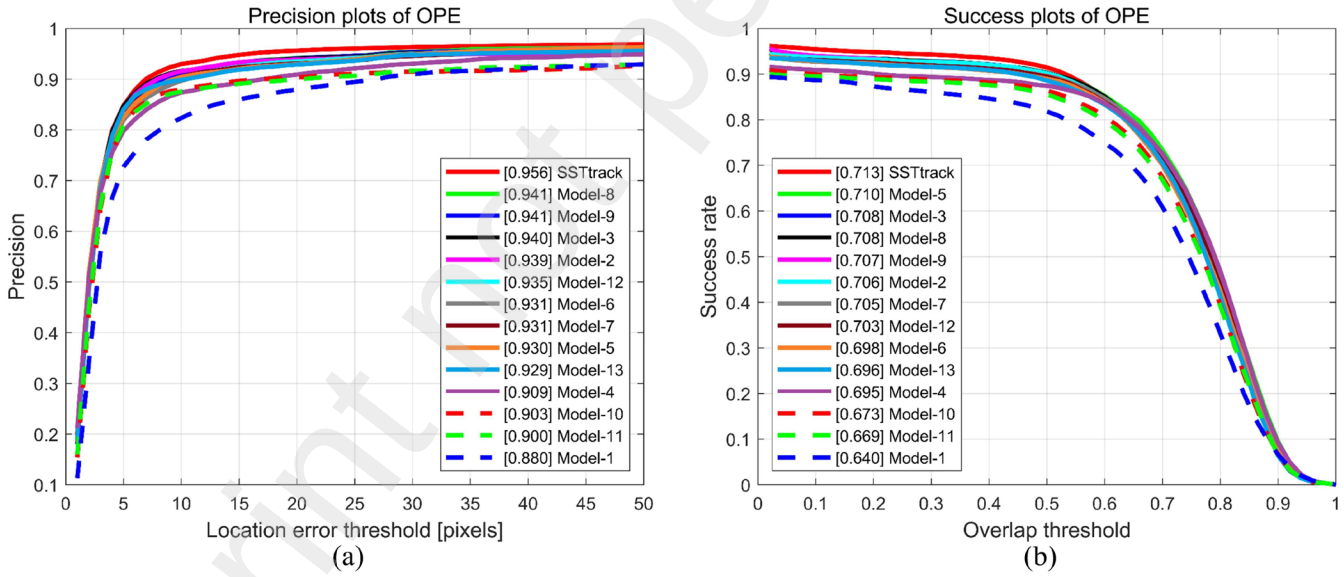

Fig. 15. The overall precision and success plots of ablation studies. (a) Precision plot. (b) Success plot.

(Yelluru Gopal et al., 2024), SiamBAG (Li et al., 2023a), AQATrack (Xie et al., 2024), SiamOHOT (Sun et al., 2023), ARCF (Huang et al., 2019), SEE-Net (Li et al., 2023b), and our SSTtrack. Benefiting from joint modeling of spectral, spatial, and temporal conditions, SSTtrack excels in mitigating interference and accurately determining object position and scale. The qualitative results affirm that SSTtrack maintains promising robustness and accuracy across various challenging scenarios. For instance, in scenarios like *excavator* with attributes IPR, SV, OPR, OCC, and DEF, and *trucker* with attributes OV, OCC, IV, and SV, our SSTtrack shows ideal effectiveness in handling these complex situations.

4.10. Ablation studies

As introduced earlier, SSTtrack mainly comprises MGA, SSA, and TCA. To assess their respective contributions, we evaluate SSTtrack alongside 13 ablation models, denoted as Model-1 to Model-13. Notably, the baseline Model-1 is only capable of performing by converting HS video to false-color video, while the other models are evaluated directly on HS video.

The overall precision and success plots of ablation studies are shown in Fig. 15. The ablation results for each component are discussed as follows.

Table 5

Ablation studies on the effect of MGA.

No.	Model	S-MGA	R-MGA	MGA	Params	Pre	Suc	ParamsI	PreI	SucI
1	Model-1	-	-	-	71.980	0.880	0.640	*n/a*	*n/a*	*n/a*
2	Model-2	✓	-	-	72.288	0.939	0.706	0.308	5.9%	6.6%
3	Model-3	-	✓	-	72.188	0.940	0.708	0.208	6.0%	6.8%
4	SSTtrack	-	-	✓	**72.555**	**0.956**	**0.713**	**0.575**	**7.6%**	**7.3%**

S-MGA denotes the use of only single-scale feature prompts in the IPEM within MGA. R-MGA denotes the replacement of the IPEM within MGA with the spectral-spatial self-expression module in SENSE. Params denotes the parameter volume (M). ParamsI indicates the improvement in the parameter volume relative to the foundation model (Model-1). The first and second scores are highlighted in bold and underlined, respectively.

Table 6

Ablation studies on the effect of IMPM within SSA

No.	Model	L-IMPM	Params	Pre	Suc	ParamsI	PreI	SucI
1	Model-1	*n/a*	71.980	0.880	0.640	n/a	n/a	n/a
2	Model-4	*null*	72.379	0.909	0.695	0.399	2.9%	5.5%
3	Model-5	0	72.396	0.930	0.710	0.416	5.0%	7.0%
4	Model-6	7	72.396	0.931	0.698	0.416	5.1%	5.8%
5	Model-7	0-3	72.449	0.931	0.705	0.469	5.1%	6.5%
6	Model-8	4-7	72.449	0.941	0.708	0.469	6.1%	6.8%
7	Model-9	0-7	72.519	0.941	0.707	0.539	6.1%	6.7%
8	Model-10	0-14	72.642	0.903	0.673	0.662	2.3%	3.3%
9	Model-11	0-19	**72.713**	0.900	0.669	**0.733**	2.0%	2.9%
10	SSTtrack	0-9	72.555	**0.956**	**0.713**	0.575	**7.6%**	**7.3%**

L-IMPM indicates the layer index of the used IMPM in the M-layer transformer encoders. *null* denotes no use of IMPM in any layer, n denotes the use of IMPM at the n-th layer, while d-m indicates that IMPM is used at layer d through layer m.

Table 7

Ablation studies on the effect of TCA.

No.	Model	TCA	V-TCA	Params	Pre	Suc	ParamsI	PreI	SucI
1	Model-1	-	-	71.980	0.880	0.640	*n/a*	*n/a*	*n/a*
2	Model-12	-	-	70.968	0.935	0.703	-1.012	5.5%	6.3%
3	Model-13	-	✓	**74.919**	0.929	0.696	**2.939**	4.9%	5.6%
4	SSTtrack	✓	-	72.555	**0.956**	**0.713**	0.575	**7.6%**	**7.3%**

V-TCA means using a self-attention layer instead of SSTFM within TCA.

4.10.1. Effect of MGA

In SSTtrack, MGA is devised to learn the multi-modal generation adaptively, comprising of the IPEM for solving the coefficient matrix and the spectral contribution prompting module for generating band contributions. To validate the effect of MGA, we conduct three ablation experiments. Model-1 is the baseline. Model-2 entails replacing the MGA in SSTtrack with S-MGA (i.e., employing only single-scale feature prompts in the IPEM within MGA). Meanwhile, Model-3 involves replacing the IPEM within MGA with R-MGA (i.e., replacing the IPEM within MGA with the spectral-spatial self-expression module in SENSE). The experimental results are presented in Table 5. Compared to Model-1, SSTtrack brings significant improvements (7.6% in Pre and 7.3% in Suc) with the introduction of only 0.575M parameters. When compared with Model-2 and Model-3, it is evident that our MGA also achieves competitive gains, i.e., 1.7% and 1.6% in the Pre score and 0.7% and 0.5% in the Suc score, respectively. These results highlight the effectiveness of MGA, enabling SSTtrack to fully leverage the rich spectral condition in the HS data.

4.10.2. Effect of IMPM within SSA

To explore a more effective and efficient SSA architecture, we embed our IMPM in a portion of the transformer encoder layer. Specifically, we establish nine embedding types. Table 6 displays experimental results, in which L-IMPM denotes the layer index of the used IMPM in the N-layer (20 layers in total) transformer encoders within HiViT (Zhang et al., 2022a). We observe that as the number of IMPM increases, the model performance steadily improves, as evident in Model-6, Model-7, Model-8, Model-9, and SSTtrack. However, further increases in the number of IMPM lead to a degradation in the magnitude of performance improvement, as observed in Model-10 and Model-11. This phenomenon may be attributed to the increased number of parameters causing the model to lose critical information about the input modality, resulting in inferior prediction results. SSTtrack achieves optimal performance with appropriate learnable parameters (0.575M). Experimental results also indicate the potential for further simplification and enhancement of the SSA architecture.

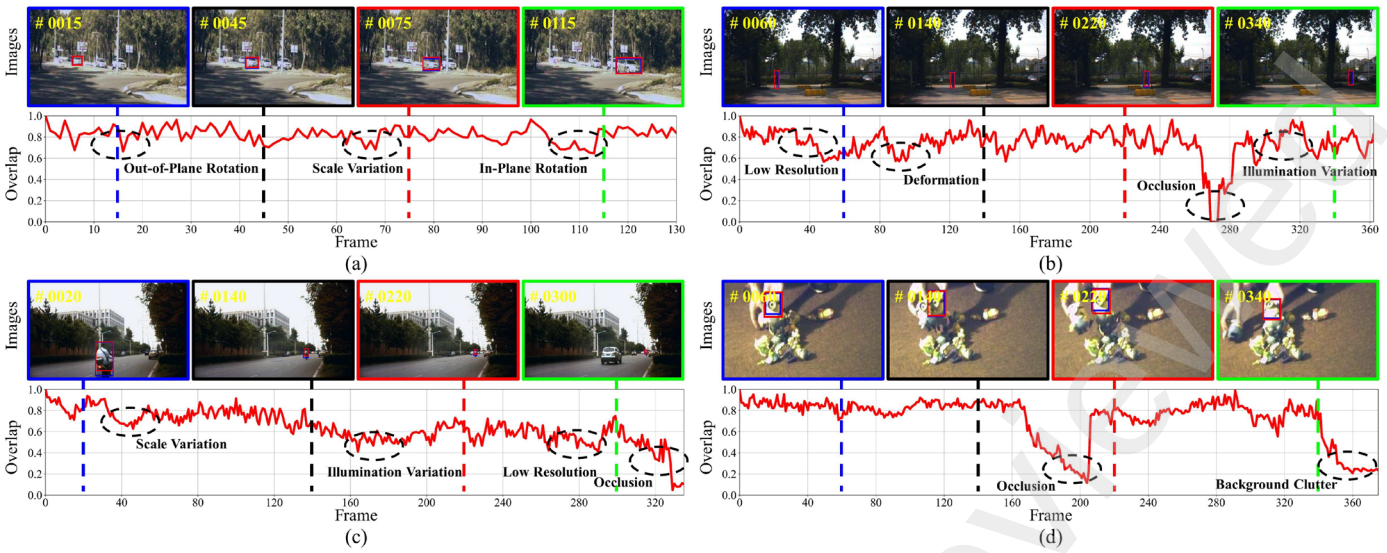

Fig. 16. Overlap curves and tracking samples obtained using SSTtrack. (a) *car2*, attribute: SV, IPR, and OPR. (b) *pedestrian2*, attribute: OCC, LR, DEF, and IV. (c) *rider1*, attribute: LR, OCC, IV, and SV. (d) *toy1*, attribute: BC and OCC. Ground truth and prediction are marked in blue and red bounding boxes, respectively.

4.10.3. Effect of TCA

TCA is devised to inject the temporal condition to guide spectral-spatial feature representations. It comprises the Transformer decoder for temporal condition transfer and the SSTFM for filtering spectral, spatial, and temporal conditions. To demonstrate the role of TCA, the ablation experiments are conducted, as depicted in Table 7. Model-12 represents the elimination of TCA from SSTtrack. Compared with Model-12, SSTtrack obtains consistent improvements by 2.1% in Pre and 1.0% in Suc. Fig. 16 visualizes the overlap curves and tracking results of SSTtrack, showcasing its ability to maintain high overlap scores even with challenging attributes.

To further validate the effect of TCA, we use a self-attention layer (Devlin et al., 2018) to replace the SSTFM within TCA for spectral-spatial-temporal condition filtering, yielding the Model-13. Experimental results reveal that SSTtrack outperforms Model-13 by 2.7% in Pre and 1.7% in Suc. In comparison to Model-12, it is observed that employing an expensive fusion strategy (Model-13) may not obtain significant improvement, or may even result in a slight degradation. It may be because the complex network introduces a higher-dimensional feature space, making it more challenging for the model to learn and generalize accurately. The SSTFM within TCA implements the filtering and fusion of spectral, spatial, and temporal conditions in an almost parameter-less manner, with results validating its performance.

5. Conclusions

This study proposes a unified HS video tracking framework (SSTtrack) via modeling spectral-spatial-temporal conditions simultaneously in an end-to-end fashion. First, a multi-modal generation adapter is designed to learn multi-modal generation and bridge the band gap. Following this, a spectral-spatial adapter is designed to dynamically transfer and interact with multiple modalities adaptively. Finally, to capture static and instantaneous object properties, we design a temporal condition adapter to inject the temporal condition for guiding spectral and spatial feature representations. The proposed framework strikes a favorable trade-off between accuracy and speed with the introduction of only 0.575M learnable parameters, making it a reliable candidate for the HS video tracking domain. Nevertheless, the current model struggles to capture longer-term spectral-spatial-temporal conditions that could enhance performance. Future work will focus on importing longer-term conditions in our framework to address the memory restrictions.

Acknowledgments

Sincerely thanks to the editors and reviewers for their valuable comments and constructive suggestions. Thanks to the related authors for providing codes and results. This work was supported by the National Natural Science Foundation of China (42230108, 423B2104, and 42271411).

References

Bertinetto, L., Valmadre, J., Golodetz, S., Miksik, O., Torr, P.H.S., 2016a. Staple: Complementary Learners for Real-Time Tracking, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1401-1409.

Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.S., 2016b. Fully-Convolutional Siamese Networks for Object Tracking, Proc. Eur. Conf. Comput. Vis. (ECCV) /IEEE Trans. Signal Process., pp. 850-865.

Bhat, G., Danelljan, M., Van Gool, L., Timofte, R., 2020. Know Your Surroundings: Exploiting Scene Information for Object Tracking. Proc. Eur. Conf. Comput. Vis. (ECCV).

Bhat, G., Danelljan, M., Van Gool, L., Timofte, R., Ieee, 2019. Learning Discriminative Model Prediction for Tracking, Ieee I Conf Comp Vis, pp. 6181-6190.

Cai, Y., Zhang, Z., Liu, X., Cai, Z., 2020. Efficient Graph Convolutional Self-Representation for Band Selection of Hyperspectral Image. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 4869-4880.

Cao, B., Guo, J., Zhu, P., Hu, Q., 2024. Bi-directional Adapter for Multi-modal Tracking. in Proc. AAAI Conf. Artif. Intell. (AAAI).

Cao, Z., Huang, Z., Pan, L., Zhang, S., Liu, Z., Fu, C., 2023. Towards Real-World Visual Tracking with Temporal Contexts. IEEE Trans. Pattern Anal. Mach. Intell.

Chen, B., Li, P., Bai, L., Qiao, L., Shen, Q., Li, B., Gan, W., Wu, W., Ouyang, W., 2022. Backbone is All Your Need: A Simplified Architecture for Visual Object Tracking. in Proc. Eur. Conf. Comput. Vis. (ECCV), 375–392.

Chen, L.L., Zhao, Y.Q., Kong, S.G., 2023a. SFA-guided mosaic transformer for tracking small objects in snapshot spectral imaging. ISPRS J. Photogramm. Remote Sens. 204, 223-236.

Chen, X., Peng, H., Wang, D., Lu, H., Hu, H., 2023b. SeqTrack: Sequence to Sequence Learning for Visual Object Tracking, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14572-14581.

Chen, Y., Yuan, Q., Tang, Y., Xiao, Y., He, J., Liu, Z., 2024a. SENSE: Hyperspectral video object tracker via fusing material and motion cues. Inf. Fusion 109, 102395.

Chen, Y.Z., Yuan, Q.Q., Tang, Y.Q., Xiao, Y., He, J., Zhang, L.P., 2024b. SPIRIT: Spectral Awareness Interaction Network With Dynamic Template for Hyperspectral Object Tracking. IEEE Trans. Geosci. Remote Sens. 62.

Chen, Z.D., Zhong, B.N., Li, G.R., Zhang, S.P., Ji, R.R., Ieee, 2020. Siamese Box Adaptive Network for Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 6667-6676.

Cui, Y., Guo, D.Y., Shao, Y.Y., Wang, Z.H., Shen, C.H., Zhang, L.Y., Chen, S.Y., 2022. Joint Classification and Regression for Visual Tracking with Fully Convolutional Siamese Networks. Int. J. Comput. Vis. 130, 550-566.

Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M., Soc, I.C., 2019. ATOM: Accurate Tracking by Overlap Maximization, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, pp. 4655-4664.

Danelljan, M., Hager, G., Khan, F.S., Felsberg, M., 2015. Learning Spatially Regularized Correlation Filters for Visual Tracking, in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp. 4310-4318.

Danelljan, M., Hager, G., Khan, F.S., Felsberg, M., 2017. Discriminative Scale Space Tracking. IEEE Trans. Pattern Anal. Mach. Intell. 39, 1561-1575.

Danelljan, M., Khan, F.S., Felsberg, M., van de Weijer, J., 2014. Adaptive Color Attributes for Real-Time Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1090-1097.

Danelljan, M., Van Gool, L., Timofte, R., Ieee, 2020. Probabilistic Regression for Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Electr Network, pp. 7181-7190.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv e-prints, arXiv:1810.04805.

Fan, H., Lin, L.T., Yang, F., Chu, P., Deng, G., Yu, S.J., Bai, H.X., Xu, Y., Liao, C.Y., Ling, H.B., Soc, I.C., 2019. LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 5369-5378.

Galoogahi, H.K., Fagg, A., Lucey, S., 2017. Learning Background-Aware Correlation Filters for Visual Tracking, Ieee I Conf Comp Vis, pp. 1144-1152.

Gao, L., Chen, L., Liu, P., Jiang, Y., Xie, W., Li, Y., 2023a. A Transformer-Based Network for Hyperspectral Object Tracking. IEEE Trans. Geosci. Remote Sens. 61.

Gao, L., Liu, P., Jiang, Y., Xie, W., Lei, J., Li, Y., Du, Q., 2023b. CBFF-Net: A New Framework for Efficient and Accurate Hyperspectral Object Tracking. IEEE Trans. Geosci. Remote Sens. 61, 1-14.

Gao, S., Zhou, C., Zhang, J., 2023c. Generalized Relation Modeling for Transformer Tracking, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 18686-18695.

Guo, D.Y., Shao, Y.Y., Cui, Y., Wang, Z.H., Zhang, L.Y., Shen, C.H., Ieee Comp, S.O.C., 2021. Graph Attention Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Electr Network, pp. 9538-9547.

Han, Y., Huang, K., 2024. ACTrack: Adding Spatio-Temporal Condition for Visual Object Tracking. arXiv e-prints, arXiv:2403.07914.

Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2012. Exploiting the Circulant Structure of Tracking-by-Detection with Kernels, Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 702-715.

Henriques, J.F., Caseiro, R., Martins, P., Batista, J., 2015. High-Speed Tracking with Kernelized Correlation Filters. IEEE Trans. Pattern Anal. Mach. Intell. 37, 583-596.

Hou, Z., Li, W., Zhou, J., Tao, R., 2022. Spatial-Spectral Weighted and Regularized Tensor Sparse Correlation Filter for Object Tracking in Hyperspectral Videos. IEEE Trans. Geosci. Remote Sens. 60.

Huang, L.H., Zhao, X., Huang, K.Q., 2021. GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild. IEEE Trans. Pattern Anal. Mach. Intell. 43, 1562-1577.

Huang, Z.Y., Fu, C.H., Li, Y.M., Lin, F.L., Lu, P., 2019. Learning Aberrance Repressed Correlation Filters for Real-Time UAV Tracking. in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2891-2900.

Javed, S., Danelljan, M., Khan, F.S., Khan, M.H., Felsberg, M., Matas, J., 2022. Visual Object Tracking with Discriminative Filters and Siamese Networks: A Survey and Outlook. IEEE Trans. Pattern Anal. Mach. Intell. PP.

Jiang, X., Wang, X., Sun, C., Zhu, Z., Zhong, Y., 2024. A Channel Adaptive Dual Siamese Network for Hyperspectral Object Tracking. IEEE Trans. Geosci. Remote Sens. 62, 1-12.

Lei, X., Cheng, W., Xu, C., Yang, W., 2024. Joint target and background temporal propagation for aerial tracking. ISPRS J. Photogramm. Remote Sens. 211, 121-134.

Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J., 2019. SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4277-4286.

Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X., 2018a. High Performance Visual Tracking with Siamese Region Proposal Network, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 8971-8980.

Li, C.L., Xue, W.L., Jia, Y.Q., Qu, Z.C., Luo, B., Tang, J., Sun, D.D., 2022. LasHeR: A Large-Scale High-Diversity Benchmark for RGBT Tracking. IEEE Trans. Image Process. 31, 392-404.

Li, F., Tian, C., Zuo, W., Zhang, L., Yang, M.H., 2018b. Learning Spatial-Temporal Regularized Correlation Filters for Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 4904-4913.

Li, W., Hou, Z.F., Zhou, J., Tao, R., 2023a. SiamBAG: Band Attention Grouping-Based Siamese Object Tracking Network for Hyperspectral Videos. IEEE Trans. Geosci. Remote Sens. 61.

Li, Y., Fu, C., Ding, F., Huang, Z., Lu, G., 2020a. AutoTrack: Towards High-Performance Visual Tracking for UAV with Automatic Spatio-Temporal Regularization, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Li, Y., Zhu, J., 2015. A Scale Adaptive Kernel Correlation Filter Tracker with Feature Integration, Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 254-265.

Li, Z., Xiong, F., Zhou, J., Lu, J., Qian, Y., 2023b. Learning a Deep Ensemble Network With Band Importance for Hyperspectral Object Tracking. IEEE Trans. Image Process. 32, 2901-2914.

Li, Z., Xiong, F., Zhou, J., Lu, J., Zhao, Z., Qian, Y., 2024. Material-Guided Multiview Fusion Network for Hyperspectral Object Tracking. IEEE Trans. Geosci. Remote Sens. 62, 1-15.

Li, Z., Xiong, F., Zhou, J., Wang, J., Lu, J., Qian, Y., 2020b. BAE-Net: A Band Attention Aware Ensemble Network for Hyperspectral Object Tracking, 2020 IEEE International Conference on Image Processing (ICIP), pp. 2106-2110.

Li, Z., Ye, X., Xiong, F., Lu, J., Zhou, J., Qian, Y., 2021. Spectral-Spatial-Temporal Attention Network for Hyperspectral Tracking, 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), pp. 1-5.

Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H., 2022. SwinTrack: A Simple and Strong Baseline for Transformer Tracking. in NeurIPS, 16,743–716,754.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection, 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2999-3007.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., Zitnick, C.L., 2014. Microsoft COCO: Common Objects in Context, Proc. Eur. Conf. Comput. Vis. (ECCV), Zurich, SWITZERLAND, pp. 740-755.

Liu, P.F., Yuan, W.Z., Fu, J.L., Jiang, Z.B., Hayashi, H., Neubig, G., 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. Acm Computing Surveys 55.

Liu, Z., Wang, X., Zhong, Y., Shu, M., Sun, C., 2022. SiamHYPER: Learning a Hyperspectral Object Tracker From an RGB-Based Tracker. IEEE Trans. Image Process. 31, 7116-7129.

Loshchilov, I., Hutter, F., 2018. Decoupled Weight Decay Regularization. ICLR, 1–19.

Lukezic, A., Vojir, T., Zajc, L.C., Matas, J., Kristan, M., 2018. Discriminative Correlation Filter Tracker with Channel and Spatial Reliability. Int. J. Comput. Vis. 126, 671-688.

Mayer, C., Danelljan, M., Pani Paudel, D., Van Gool, L., 2021. Learning Target Candidate Association to Keep Track of What Not to Track. in Proc. IEEE Int. Conf. Comput. Vis. (ICCV).

Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B., 2018. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild, 15th European Conference on Computer Vision (ECCV), Munich, GERMANY, pp. 310-327.

Ouyang, E., Wu, J., Li, B., Zhao, L., Hu, W., 2022. Band Regrouping and Response-Level Fusion for End-to-End Hyperspectral Object Tracking. IEEE Geosci. Remote Sens. Lett. 19.

Possegger, H., Mauthner, T., Bischof, H., Ieee, 2015. In Defense of Color-based Model-free Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 2113-2120.

Qian, K., Zhou, J., Xiong, F., Zhou, H., Du, J., 2018. Object Tracking in Hyperspectral Videos with Convolutional Features and Kernelized Correlation Filter, International Conference on Smart Multimedia., pp. 308-319.

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S., 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 658-666.

Shi, L., Zhong, B., Liang, Q., Li, N., Zhang, S., Li, X., 2024. Explicit Visual Prompts for Visual Object Tracking. arXiv e-prints, arXiv:2401.03142.

Song, Y.B., Ma, C., Wu, X.H., Gong, L.J., Bao, L.C., Zuo, W.M., Shen, C.H., Lau, R.W.H., Yang, M.H., 2018. VITAL: Visual Tracking via Adversarial Learning, Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 8990-8999.

Su, Y., Mei, S., Zhang, G., Wang, Y., He, M., Du, Q., Ieee, 2022. Gaussian Information Entropy Based Band Reduction for Unsupervised Hyperspectral Video Tracking, IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 791-794.

Sun, C., Wang, X., Liu, Z., Wan, Y., Zhang, L., Zhong, Y., 2023. SiamOHOT: A Lightweight Dual Siamese Network for Onboard Hyperspectral Object Tracking via Joint Spatial-Spectral Knowledge Distillation. IEEE Trans. Geosci. Remote Sens., 1-1.

Tang, Y., Huang, H., Liu, Y., Li, Y., 2023. A Siamese network-based tracking framework for hyperspectral video. Neural Comput. Appl. 35, 2381-2397.

Tang, Y., Liu, Y., Huang, H., 2022a. Target-aware and spatial-spectral discriminant feature joint correlation filters for hyperspectral video object tracking. Comput. Vis. Image Underst. 223.

Tang, Y., Liu, Y., Ji, L., Huang, H., 2022b. Robust Hyperspectral Object Tracking by Exploiting Background-Aware Spectral Information With Band Selection Network. IEEE Geosci. Remote Sens. Lett. 19.

Uzkent, B., Rangnekar, A., Hoffman, M.J., 2019. Tracking in Aerial Hyperspectral Videos Using Deep Kernelized Correlation Filters. IEEE Trans. Geosci. Remote Sens. 57, 449-461.

Wang, H., Liu, X., Li, Y., Sun, M., Yuan, D., Liu, J., 2024. Temporal Adaptive RGBT Tracking with Modality Prompt. arXiv e-prints, arXiv:2401.01244.

Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.S., 2019. Fast Online Object Tracking and Segmentation: A Unifying Approach, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 1328-1338.

Wang, S., Qian, K., Chen, P., Ieee, 2022. BS-SiamRPN: Hyperspectral Video Tracking based on Band Selection and the Siamese Region Proposal Network, 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS).

Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y., 2023. Autoregressive Visual Tracking, 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9697-9706.

Wu, Y., Lim, J., Yang, M.H., 2015. Object Tracking Benchmark. IEEE Trans. Pattern Anal. Mach. Intell. 37, 1834-1848.

Xie, F., Wang, C., Wang, G., Cao, Y., Yang, W., Zeng, W., 2022. Correlation-Aware Deep Tracking, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 8741-8750.

Xie, J., Zhong, B., Mo, Z., Zhang, S., Shi, L., Song, S., Ji, R., 2024. Autoregressive Queries for Adaptive Tracking with Spatio-Temporal Transformers. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Xiong, F., Zhou, J., Qian, Y., 2020. Material Based Object Tracking in Hyperspectral Videos. IEEE Trans. Image Process. 29, 3719-3733.

Xu, Y.D., Wang, Z.Y., Li, Z.X., Yuan, Y., Yu, G., Assoc Advancement Artificial, I., 2020. SiamFC plus plus : Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines, in Proc. AAAI Conf. Artif. Intell. (AAAI), pp. 12549-12556.

Xuefeng Zhu, Tianyang Xu, Zongtao Liu, Zhangyong Tang, Xiaojun Wu, Kittler, J., 2024. UniMod1K: Towards a More Universal Large-Scale Dataset and Benchmark for Multi-modal Learning. Int. J. Comput. Vis., 1-16.

Yan, B., Peng, H., Fu, J., Wang, D., Lu, H., 2021a. Learning Spatio-Temporal Transformer for Visual Tracking. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Yan, B., Peng, H., Wu, K., Wang, D., Fu, J., Lu, H., 2021b. LightTrack: Finding Lightweight Neural Networks for Object Tracking via One-Shot Architecture Search, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 15175-15184.

Yang, J., Li, Z., Zheng, F., Leonardis, A., Song, J., 2022. Prompting for Multi-Modal Tracking.

Ye, B., Chang, H., Ma, B., Shan, S., 2022. Joint Feature Learning and Relation Modeling for Tracking: A One-Stream Framework. in Proc. Eur. Conf. Comput. Vis. (ECCV).

Yelluru Gopal, G., Amer, M.A., 2024. Separable Self and Mixed Attention Transformers for Efficient Object Tracking. in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 6708-6717.

Zhang, L.C., Gonzalez-Garcia, A., Van De Weijer, J., Danelljan, M., Khan, F.S., Ieee, 2019a. Learning the Model Update for Siamese Trackers, Ieee I Conf Comp Vis, Seoul, SOUTH KOREA, pp. 4009-4018.

Zhang, X., Tian, Y., Huang, W., Ye, Q., Dai, Q., Xie, L., Tian, Q., 2022a. HiViT: Hierarchical Vision Transformer Meets Masked Image Modeling. arXiv e-prints, arXiv:2205.14949.

Zhang, Y., Li, X., Wang, F., Wei, B., Li, L., Ieee, 2022b. A Fast Hyperspectral Object Tracking Method Based On Channel Selection Strategy, 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS).

Zhang, Z., Qian, K., Du, J., Zhou, H., 2021. Multi-Features Integration Based Hyperspectral Videos Tracker, 2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), pp. 1-5.

Zhang, Z.P., Peng, H.W., Soc, I.C., 2019b. Deeper and Wider Siamese Networks for Real-Time Visual Tracking, Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, pp. 4586-4595.

Zhao, C., Liu, H., Su, N., Yan, Y., 2022. TFTN: A Transformer-Based Fusion Tracking Framework of Hyperspectral and RGB. IEEE Trans. Geosci. Remote Sens. 60.

Zhu, J., Lai, S., Chen, X., Wang, D., Lu, H., 2023a. Visual Prompt Multi-Modal Tracking. in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR).

Zhu, X.G., Zhang, H.R., Hu, B., Huang, K.P., Arun, P.V., Jia, X.P., Zhao, D., Wang, Q., Zhou, H.X., Yang, S.W., 2023b. DSP-Net: A Dynamic Spectral–Spatial Joint Perception Network for Hyperspectral Target Tracking. IEEE Geosci. Remote Sens. Lett. 20.

Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W., 2018. Distractor-Aware Siamese Networks for Visual Object Tracking, Proc. Eur. Conf. Comput. Vis. (ECCV), pp. 103-119.