PHTrack: Prompting for Hyperspectral Video Tracking

Yuzeng Chen, Yuqi Tang, *Member*, *IEEE*, Xin Su, *Member*, *IEEE*, Jie Li, *Member*, *IEEE*, Yi Xiao, *Graduate Student Member*, *IEEE*, Jiang He, *Graduate Student Member*, *IEEE*, and Qiangqiang Yuan, *Member*, *IEEE*

***Abstract*—Hyperspectral (HS) video captures continuous spectral information of objects, enhancing material identification in tracking tasks. It is expected to overcome the inherent limitations of RGB and multi-modal tracking, such as finite spectral cues and cumbersome modality alignment. However, HS tracking faces challenges like data anxiety, band gaps, and huge volumes. In this study, inspired by prompt learning in language models, we propose the Prompting for Hyperspectral Video Tracking (PHTrack) framework. PHTrack learns prompts to adapt foundation models, mitigating data anxiety and enhancing performance and efficiency. First, the modality prompter (MOP) is proposed to capture rich spectral cues and bridge band gaps for improved model adaptation and knowledge enhancement. Additionally, the distillation prompter (DIP) is developed to refine cross-modal features. PHTrack follows feature-level fusion, effectively managing huge volumes compared to traditional decision-level fusion fashions. Extensive experiments validate the proposed framework, offering valuable insights for future research. The code and data will be available at https://github.com/YZCU/PHTrack.**

***Index Terms*—Hyperspectral video tracking, Prompt learning, Self-expression model, Material information.**

I. Introduction

Visual object tracking aims to establish the association of the object in a video, which finds practical applications in various fields, such as video surveillance, human-machine interaction, and medical imaging [1]. Significant progress has been made using the red-green-blue (RGB) modality [2], [3], [4]. There are still challenges to overcome in complex situations like similar appearance, low light, poor visibility, and background clutter due to the limited spectral cues [5]. To address these challenges, multi-modal data with complementary cues has been uncovered for enhanced performance, such as RGB plus thermal infrared (RGB-T) [6], RGB plus depth (RGB-D) [7], and RGB plus event (RGB-E) [8], as shown in Fig. 1. However, the routine multi-modal object tracking mission involves the multi-device imaging. In RGB-T tracking, for example, the charge-coupled device and an infrared camera are usually mounted on a platform to simultaneously record multi-modal data [9]. Despite the proximity, multi-device imaging makes it difficult to capture the same scene, especially for small objects at long distances. Hence, the modal alignment has become a standard practice [9], although it may lead to image distortion problems, as shown in Fig. 1(a) and (b).

Breakthroughs in imaging technology have led to the development of hyperspectral (HS) cameras, which are powerful tools for capturing the continuous spectral characteristics for object detection [10], [11], classification [12], image enhancement [13], and change detection [14]. The camera enables trackers to identify materials based on their spectral signatures [15]. Notably, the HS data records the material radiated signals in distinct bands that are shot from the same viewpoint. Indeed, conventional multi-modal devices capture information in different radiated bands (e.g., RGB and infrared) or different mechanisms (e.g., depth and event). Integrating robust multi-modal tracking into HS video tracking holds promise for overcoming spectral limitations and alignment issues in RGB and multi-modal tracking domains.

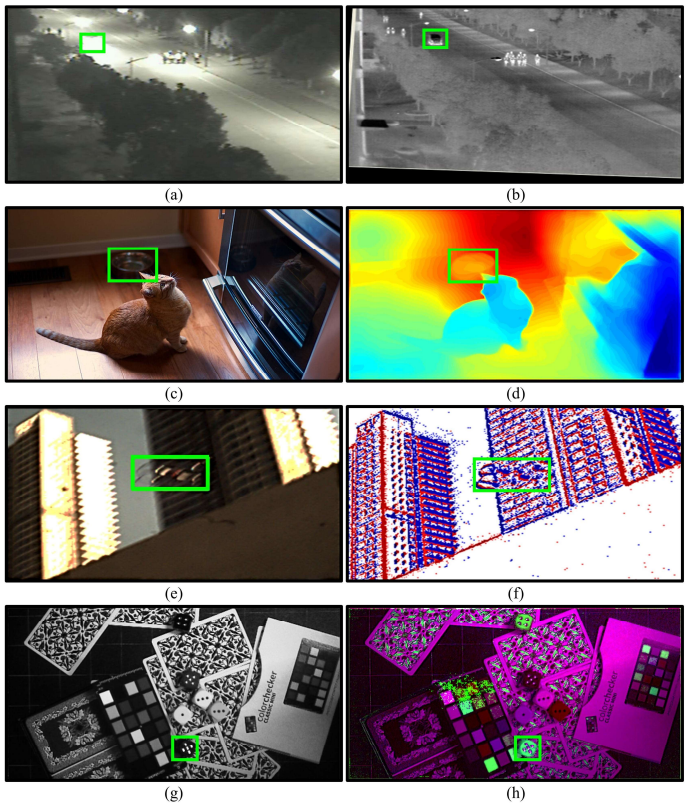


Fig. 1. Sample of multi-modal data. (a) and (b) show RGB and Thermal infrared modalities. (c) and (d) are RGB and Depth modalities. (e) and (f) are RGB and Event modalities. (g) and (h) are multi-modal data generated from the HS modality.

HS trackers mainly rely on discriminative features and generalized models. Previous studies, such as TASSCF [16], TSCFW [17], and MHT [18] inherit the correlation filter using

This work was supported in part by the National Natural Science Foundation of China under Grant 42230108. *(Corresponding author: Xin Su, Jie Li)*

Yuzeng Chen, Jie Li, Yi Xiao, Jiang He, and Qiangqiang Yuan are with the School of Geodesy and Geomatics, Wuhan University, Wuhan 430079, China (e-mail: yuzeng_chen@whu.edu.cn; yqiang86@gmail.com).

Xin Su is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: xinsu.rs@whu.edu.cn).

Yuqi Tang is with the School of Geosciences and Info-Physics, Central South University, Changsha 410012, China (e-mail: yqtang@csu.edu.cn).

hand-crafted features. However, the limited discrimination of hand-crafted features can affect tracking performance [19]. To address this problem, recent state-of-the-art (SOTA) works, such as SiamBAG [20], SPIRIT [15], SENSE [21], and SiamOHOT [22], primarily use the Siamese network embedding deep features for robust object representations. Discriminative features are fundamental, and a robust model determines the ceiling of performance [23]. As evidenced in RGB tracking, traditional correlation filters struggle to achieve comparable results against Siamese networks [24], [25].

Currently, the field of HS video tracking is encountering several challenges. First, data anxiety, i.e., the scarcity and poor transferability of HS video datasets, hinders the direct training of a generalized HS tracking model [5], [15]. Second, the discrepancy in band numbers between RGB and HS modalities, known as band gaps, poses a challenge [26]. Third, the huge volumes where the computational cost of processing HS data is expensive due to dense bands leads to slower operations [22]. RGB tracking benefits from extensive datasets like TrackingNet [27] and ImageNet [28], fostering well-trained foundational models. For the data anxiety, one conceivable idea is to use the pre-trained RGB foundation model and perform full fine-tuning on HS training sets. While effective, the risk of overfitting is increasing due to the scarcity of large-scale HS video datasets, contrasting with the data-intensive demands of data-driven models. Additionally, full fine-tuning is time-consuming, and parameter storage imposes a significant burden [29]. Addressing these data constraints is crucial for achieving robust HS models. Prompt tuning, effective in natural language processing, enhances foundation models by injecting textual prompts. Regarding band gaps, approaches like SiamHT [30], BAHT [31], and BS-SiamRPN [32], attempt to convert HS images into a three-channel representation through manual selection and dimension reduction, which inevitably leads to loss and distortion of critical material cues [26]. In contrast, methods such as MHT [18], BRRF-Net [33], SiamBAG [20], SEE-Net [26], and SST-Net [19], aim to capitalize on rich spectral information, yielding competitive results. Despite their potential, these methods often rely on decision-level fusion, potentially limiting efficiency by producing multiple outputs (e.g., response maps and bounding boxes) per inference session [15].

Motivated by the above analysis, this article proposes the Prompting for Hyperspectral Video Tracking (PHTrack) framework. Leveraging shared knowledge in feature extraction and attentional patterns between HS and RGB modalities, we address data scarcity in HS tracking through prompt learning with prompt tuning using the RGB-based foundation model. PHTrack includes trainable prompters: the modality prompter (MOP) learns multi-modal generation from HS images to extract spectral cues and bridge HS-RGB domains, while the distillation prompter (DIP) integrates cross-modal features by refining adjacent modality information, enhancing classification and regression networks of the foundation model. PHTrack's feature-level fusion strategy alleviates the challenge of processing huge volumes compared to traditional decision-level fusion methods. Major contributions can be outlined as follows.

1) We propose the prompt tracking framework, PHTrack, which harnesses learned prompts to effectively adapt an off-the-shelf foundation model from the RGB domain to the HS tracking domain, addressing the data anxiety.
2) A modality prompter is developed to learn the multi-modal generation of HS images, enabling the extraction of rich spectral cues and bridging band gaps between HS and RGB domains to stimulate prior knowledge.
3) A distillation prompter is designed to integrate cross-modal complementary features, notably alleviating the challenge of processing large volumes of HS modality compared to decision-level fusion methods.

Extensive experiments are conducted to validate the proposed method. The remainder is organized as follows. Section II provides related work. In Section III, we detail the proposed method. Section IV presents the experimental results and analysis. Lastly, Section V formulates the conclusions and highlights the main contribution.

II. Related Work

A. HS Video Tracking

Generative and discriminative paradigms are used in HS video tracking. In the early stages, researchers focus on the generative paradigm, which involves creating a model to represent the object and finding similar regions [34]. However, recent HS trackers have predominantly inherited the discriminative paradigm including correlation filters and Siamese networks. Some HS trackers, such as CNHT [35], MHT [18], TSCFW [17], TASSCF [16], and MFI [36], are modeled on correlation filters to make use of full band information. Despite incorporating hand-crafted and/or deep features, these correlation filter-based HS trackers achieve limited success [26]. As mentioned earlier, reliable tracking relies on discriminative features, while the performance ceiling is determined by the robustness of the model. The simplicity and discriminative capabilities of the Siamese network have garnered significant attention. SOTA HS trackers, such as SiamOHOT [22], SEE-Net [26], SiamBAG [20], SiamHYPER [5], CBFF-Net [37], SiamHT [30], BRRF-Net [33], SSTtrack [38], and Trans-DAT [39], have integrated Siamese networks to deliver exceptional performance, laying a solid foundation for further research. For example, SiamBAG [20] develops a novel Siamese framework using band attention grouping to address insufficient training data challenges. CBFF-Net [37] develops a bidirectional multiple deep feature fusion module and a cross-band group attention module to enhance interaction information among HS bands. Extensive experiments confirm their effectiveness. Nevertheless, issues like data anxiety, band gaps, and huge volumes inherently impede the robustness of HS video object tracking. Towards this end, we propose PHTrack, a streamlined framework integrating prompt learning to maximize the scalability of foundation models.

B. Visual Prompt Learning

In recent years, prompt learning has significantly enhanced performance across various natural language processing tasks [40]. Interestingly, this approach has also proven to be effective in visual tasks. For example, AdaptFormer [41] fine-tunes lightweight modules to adapt the pre-trained backbone for scalable vision recognition tasks. VPT [42] develops an effective alternative to prompt tuning the large-scale Transformer models. EVP [43] emphasizes learnable parameters that accentuate

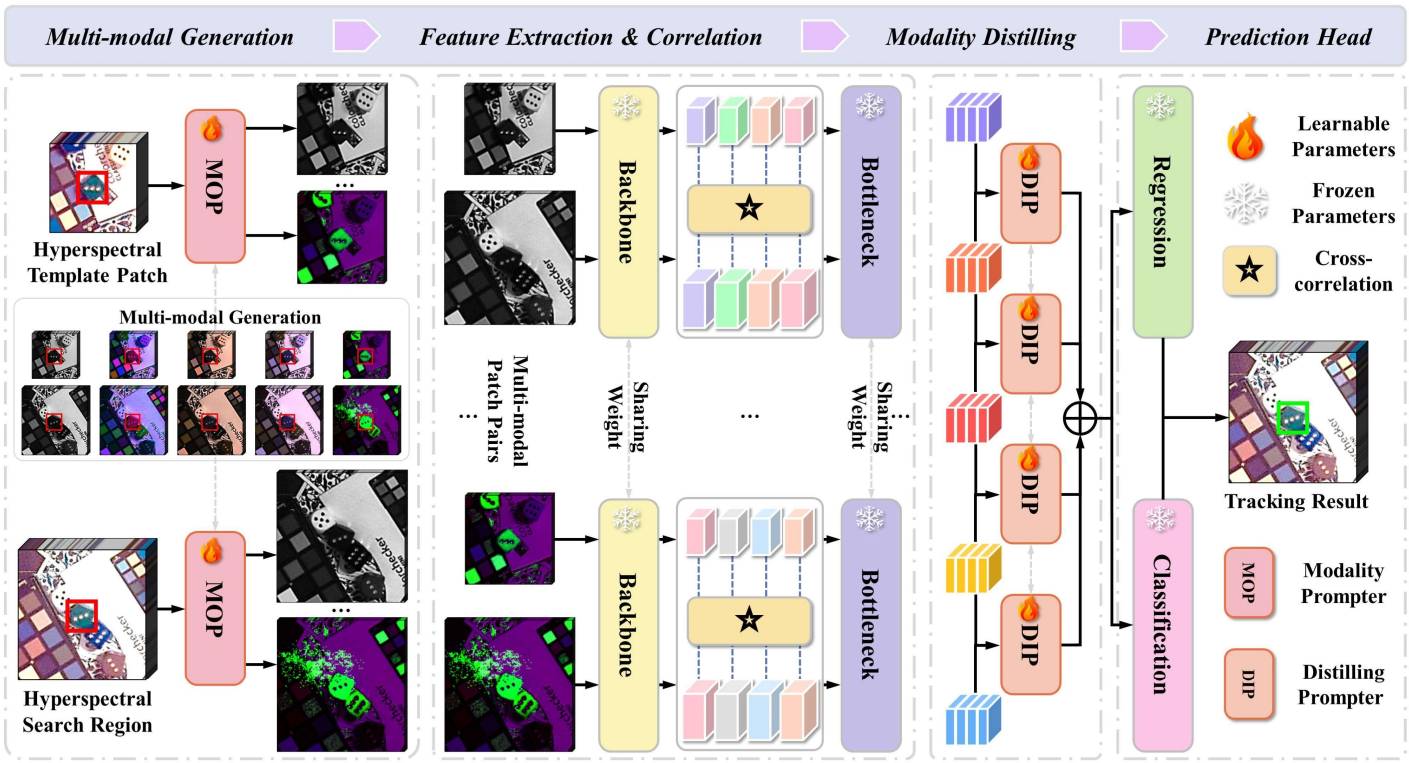


Fig. 2. Overview of the proposed HS video object tracking framework: PHTrack.

explicit visual content, incorporating features from frozen patch embedding and high-frequency input components. Recently, researchers have introduced visual prompt tuning for generative transfer learning [44]. Extensive experiments demonstrate the effectiveness of knowledge transfer, yielding significantly improved image generation quality. In [45], researchers embed prompt learning to harness domain-specific knowledge from specialized foundation models. They employ a quaternion network to transfer the strong recognition capability of vision-language models across different domains. In tracking, ProTrack [46] innovatively converts multi-modal inputs into a single modality through the prompt paradigm, which enables efficient capabilities acquired from pre-trained RGB foundation models. ViPT [29] applies prompt learning to multi-modal tracking by tailoring prompts to each modality. MPTrack [47] enables efficient multi-template aggregation and interaction with relevant queries and clues for visual object tracking. BAT [48] employs a bi-directional adapter to enhance multi-modal tracking, dynamically analyzing evolving dominant-auxiliary relationships among modalities and extracting valuable information from the pre-trained foundation model. Considerable research has demonstrated the effectiveness and generalization of prompt learning in the visual community.

III. Proposed Approach

A. Overview

The PHTrack framework consists of MOP and DIP with a pre-trained foundation network comprising backbone, bottleneck, regression, and classification sub-networks (Fig. 2). Initially, HS images of arbitrary scales are cropped into patches (i.e., search and template patches), and input to the MOP for generating multi-modal data to extract spectral information, which is then utilized for feature extraction and correlation. DIP integrates cross-modal features by modality distillation, and object state estimation is handled by the prediction head. PHTrack minimizes parameter burden by sharing weights across components.

B. Modality Prompter

As discussed, the HS modality offers rich spectral information, addressing RGB limitations. However, overfitting risks arise due to the lack of large-scale HS datasets and the data demands of data-driven models, hindering the direct learning of a generalized HS model. Thus, we explore prompt learning to leverage RGB-based foundation models. Specifically, we propose MOP to extract rich spectral cues and bridge band gaps, leveraging prior knowledge. MOP inherits the learning-to-optimize fashion to derive the HS self-expression model and generate multi-modal data from HS images. Its key components include HS self-expression modeling and multi-modal generation.

1) HS Self-expression Modeling: HS self-expression modeling involves selecting informative bands from an original set. Each band is reconstructed using a self-expression matrix $C \in \mathbb{R}^{B\times B}$, where B is the number of bands. Within C, the j-th column, i-th row, and (i,j)-th elements are denoted as c_j, c^i, and c_{ij}, respectively. C reveals intrinsic relationships among spectral bands and can be used to generate multi-modal data.

Given an HS video, each frame X can be represented as $X = [x_1, x_2, \dots, x_B] \in \mathbb{R}^{D\times B}$, where $D = M \times N$ is the number of pixels, and $x_i \in \mathbb{R}^{M\times N}$ is the i-th band vector. The HS self-expression model for the band set X is represented as:

$$argmin\|C\|_{1,2}, \text{s.t.}, X = XC + E, diag(C) = 0, C \geq 0 \quad (1)$$

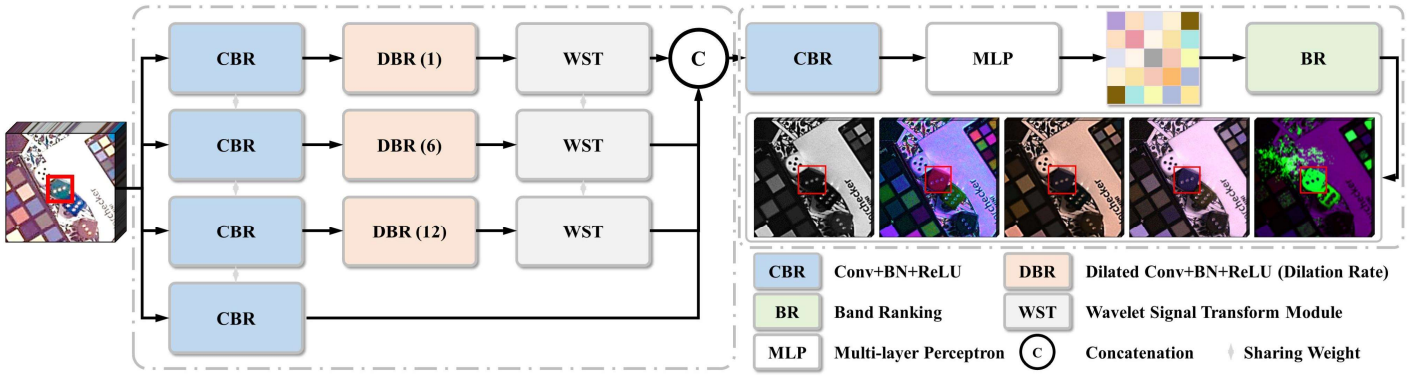


Fig. 3. Architecture of the proposed MOP.

$$\|C\|_{1,2} = \sum_{i=1}^{B} \left\|c^i\right\|_2, \sum_{i=1}^{B} c_{ij} = 1, \forall j. \tag{2}$$

where $E \in \mathbb{R}^{D\times B}$ denotes the residual induced by Gaussian noise. $diag(C) = 0$ aims to prevent the trivial solution. $C \geq 0$ ensures that each element of c_j indicates the probability of representing x_j. $\|C\|_{1,2}$ is the sum of l_2-norm of all row vectors c^i.

2) Multi-modal Generation: The key to HS self-expression modeling lies in solving for C. The traditional solution usually involves a time-consuming iterative optimization strategy. However, MOP inherits the learning-to-optimize strategy, enabling the efficient solution of C and facilitating modality generation from the HS image. MOP (Fig. 3) comprises multi-scale feature extraction and modality generation components.

To achieve multi-scale feature extraction, our method fosters information interaction at different scales to mine rich contextual representations. It is built upon the Atrous Spatial Pyramid Pooling [49] and comprises four branches. The first three branches employ dilated convolutions with dilation rates of 1, 6, and 12 to expand the receptive field and capture surrounding information. The process is formulated as follows:

$$\begin{cases} Ms_1 = DBR^1\big(CBR(X)\big), \\ Ms_2 = DBR^6\big(CBR(X)\big), \\ Ms_3 = DBR^{12}\big(CBR(X)\big), \\ Ms_4 = CBR(X), \end{cases} \tag{3}$$

where Ms_i denotes the multi-scale feature extracted from the i-th branch, CBR signifies the standard convolution followed by batch normalization (BN) and Rectified Linear Unit (ReLU). DBR^q denotes dilated convolution with a dilation rate of q followed by BN and ReLU. Next, the multi-scale feature branch aligns cross-scale information using the proposed wavelet signal transform (WST) module. It is an indisputable fact that the cosine and sine functions are the most versatile wavelets in signal transformation and signal classification, which are expected to capture the signal of objects at variable sizes or salience, providing effective contextual information [50]. The mathematical expression for the WST module (Fig. 4) is

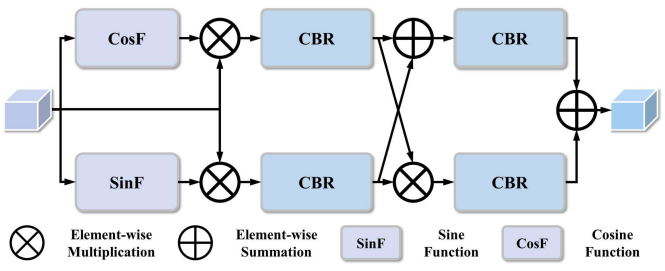


Fig. 4. Architecture of the proposed WST module.

$$\begin{cases} Ms_i^c = CBR\big(Ms_i \otimes Cos(Ms_i)\big), \\ Ms_i^s = CBR\big(Ms_i \otimes Sin(Ms_i)\big), \\ Ms_i^{mul} = Ms_i^c \otimes Ms_i^s, \\ Ms_i^{sum} = Ms_i^c \oplus Ms_i^s, \\ Ms_i^{wst} = CBR\big(Ms_i^{mul}\big) \oplus CBR(Ms_i^{sum}), \end{cases} \tag{4}$$

where $\otimes$ and $\oplus$ denote the element-wise multiplication and summation. Ms_i^{wst} is the output of the WST module. These features, generated at four scales, are then concatenated and further passed by a bottleneck layer to adjust the channels. Inspired by the efficiency and semantic interaction of the multi-layer perceptron (MLP), these features are decoded by an MLP to obtain C. To summarize, the process is described as:

$$\begin{cases} Ms^{cat} = Cat(Ms_1^{wst}, Ms_2^{wst}, Ms_3^{wst}, Ms_4), \\ H = MLP\big(CBR(Ms^{cat})\big), \\ C = H^T H \end{cases} \tag{5}$$

where Cat is concatenation. MLP means MLP. H is the attention matrix of Ms^{cat}.

For the multi-modal generation, $C \in \mathbb{R}^{B\times B}$ is a self-expression coefficient matrix. c_j denotes the representation coefficient of the j-th band represented by all the remaining bands (including itself), while c^i denotes the contribution of the i-th band to the spectral reconstruction. The more important a band is, the larger the contribution. By aggregating contributions from C, we derive rankings that indicate the significance of each band, which are subsequently used for generating multi-modal data. More concretely, C is normalized along the column direction by $\check{c}_j = |c_{ij}|/\|c_j\|_2, \forall i$, followed by $z_i = \left\|\check{c}^i\right\|_1$, where $\check{c}_j$ is the normalization output of the j-th column of C. $\check{c}^i$ stands for the i-th row of the normalized C. The cumulative contribution of B bands is denoted as $Z = [z_1, z_2, \cdots, z_B] \in \mathbb{R}^{B\times 1}$. Then, we sort all HS bands in descending order and group them to get multi-modal data $Q = [q_1, q_2, \cdots, q_K]$, where $K = B//3$ and $q_i \in \mathbb{R}^{M\times N\times 3}$.

C. Distilling Prompter

Guided by MOP principles, PHTrack learns the multi-modal generation, effectively bridging band gaps. These data are then

used for feature extraction and correlation. Neighboring modalities often provide rich information for object representations. Multi-modal features encompass discriminative information from other modalities alongside the present one. However, primitive fusion strategies, such as concatenation and element-wise summation, often fail to consider the varying importance of multi-modal features across different spatial and channel contexts. Different channel and spatial features vary in their object representation capabilities, and treating them equally can impact tracking performance. Motivated by this, we propose DIP to integrate cross-modal complementary features by refining adjacent modalities, thereby enhancing the classification and regression networks of the foundation model. Weight parameters are shared among DIPs to adhere to the prompt learning paradigm, minimizing parameter count.

Fig. 5 depicts the DIP architecture, which fuses two input features using cross-attention and channel-spatial re-weighting. Multi-head attention, a critical component of the Transformer, comprises several parallel non-local attention layers. It aims to capture global dependencies across all positions of the input. Specifically, we compute the Query, Key, and Value by:

$$\begin{cases} \mathcal{Q}_t = RT\big(CBR(F_t^0)\big), \\ \mathcal{K}_t = R\big(CBR(F_t^0)\big), \\ \mathcal{V}_t = RT\big(CBR(F_t^0)\big), \end{cases} \tag{6}$$

$$\begin{cases} \mathcal{Q}_b = RT\big(CBR(F_b^0)\big), \\ \mathcal{K}_b = R\big(CBR(F_b^0)\big), \\ \mathcal{V}_b = RT\big(CBR(F_b^0)\big), \end{cases} \tag{7}$$

where RT denotes the Reshape & Transpose. R denotes the Reshape. F_t^0 and F_b^0 are the input features from the top and bottom, respectively. Matrices $\mathcal{Q}$, $\mathcal{K}$, and $\mathcal{V}$ denote query, key of dimension d_k, and value, respectively. The remainder of the DIP is formulated by:

$$\begin{cases} F_t^1 = CBR\left(R\left(SF\left(\dfrac{\mathcal{Q}_b\mathcal{K}_t{}^T}{\sqrt{d_k}}\right)\mathcal{V}_t\right)\right), \\ F_b^1 = CBR\left(R\left(SF\left(\dfrac{\mathcal{Q}_t\mathcal{K}_b{}^T}{\sqrt{d_k}}\right)\mathcal{V}_b\right)\right), \\ W = CBR\big(Cat(F_t^1, F_b^1)\big), \\ F = W \otimes F_t^0 \oplus (1-W) \otimes F_b^0, \end{cases} \tag{8}$$

where W is the weight. F is the result of DIP.

D. Loss Function

Here we present the multi-task loss $\mathcal{L}_{total}$:

$$L_{total} = \mu_1 L_{rec} + \mu_2 L_{cls} + \mu_3 L_{reg} + \mu_4 L_{cen}, \tag{9}$$

where L_{rec}, L_{cls}, L_{reg}, and L_{cen} represent the reconstruction, classification, regression, and center-ness losses, respectively. μ_1, μ_2, μ_3 and μ_4 denote corresponding weights. During training, we empirically set μ_1, μ_2, μ_3, and μ_4 to 3.0, 1.0, 3.0, and 1.0 for all experiments. Concretely, L_{rec} is the average L_1 loss of the template and the search patches. L_{cls} is the cross-entropy loss for classification. L_{reg} is the IoU loss between the estimation and ground truth, defined as:

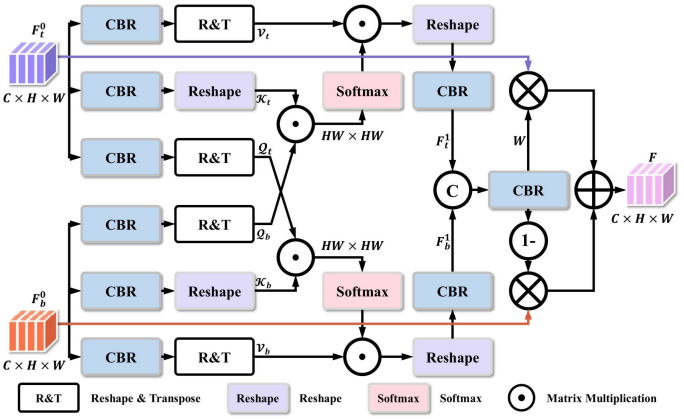


Fig. 5. Architecture of the proposed DIP.

$$\begin{cases} \hat{v}_{(i,j)}^l = \hat{l} = p_x - p_x^{lt}, \\ \hat{v}_{(i,j)}^t = \hat{t} = p_y - p_y^{lt}, \\ \hat{v}_{(i,j)}^r = \hat{r} = p_x^{rb} - p_x, \\ \hat{v}_{(i,j)}^b = \hat{b} = p_y^{rb} - p_y, \end{cases} \tag{10}$$

$$\begin{cases} \mathbb{I}\big(\hat{v}_{(i,j)}\big) = \begin{cases} 1 & if\ \forall(\hat{l},\hat{t},\hat{r},\hat{b}) > 0 \\ 0 & otherwise \end{cases}, \\ L_{reg} = \dfrac{1}{\sum \mathbb{I}\big(\hat{v}_{(i,j)}\big)} \sum_{i=1}^{w} \sum_{j=1}^{h} \mathbb{I}\big(\hat{v}_{(i,j)}\big) L_{iou}\big(A_{reg}(i,j), \hat{v}_{(i,j)}\big), \end{cases} \tag{11}$$

where $\hat{\cdot}$ is the estimation item. $\hat{l}$, $\hat{t}$, $\hat{r}$, and $\hat{b}$ indicate the estimated distance from the corresponding position to four sides (i.e., left, top, right, and bottom) of the box in the search region. (p_x, p_y) is the corresponding positions of points (i,j). (p_x^{lt}, p_y^{lt}) and (p_x^{rb}, p_y^{rb}) are the left-top and right-bottom corners of ground truth. $\hat{o}_{(i,j)}$ is the regression object at the regression response map $A_{reg} \in \mathbb{R}^{w \times h \times 4}$, $\mathcal{L}_{iou}$ is the IoU loss.

The object's center can impact prediction, with a decrease in bounding box quality as the predicted center deviates from the ground truth. To mitigate the concern, L_{cen} is defined as:

$$\begin{cases} S(i,j) = \mathbb{I}\big(\hat{v}_{(i,j)}\big) * \sqrt{\dfrac{min(\hat{t},\hat{b})}{max(\hat{t},\hat{b})} \times \dfrac{min(\hat{l},\hat{r})}{max(\hat{l},\hat{r})}}, \\ L_{cen} = \dfrac{-1}{\sum \mathbb{I}\big(\hat{v}_{(i,j)}\big)} \sum_{\mathbb{I}\left(\hat{v}_{(i,j)}\right)==1} S(i,j) * log\, A_{cen}(i,j) \\ \qquad +\big(1 - S(i,j)\big) * log\big(1 - A_{cen}(i,j)\big), \end{cases} \tag{12}$$

where $S(i,j)$ records the center-ness score for each point (i,j) of the center-ness response map $A_{cen} \in \mathbb{R}^{w \times h \times 1}$ in the search patch. $S(i,j)$ is up to 1 while satisfying $\hat{t} = \hat{b}$, $\hat{l} = \hat{r}$ and $\mathbb{I}\big(\hat{v}_{(i,j)}\big) = 1$. In this way, the predicted center would be drawn as close to the ground truth as possible.

IV. Experiments

A. Experimental Setups

1) Datasets: PHTrack undergoes training and testing using the HS dataset from the Hyperspectral Object Tracking Competition (HOTC) [18], which comprises 40 sets for training and 35 sets for testing. Each dataset with frame-level annotations contains three video types: HS video (16 bands), false color video (3

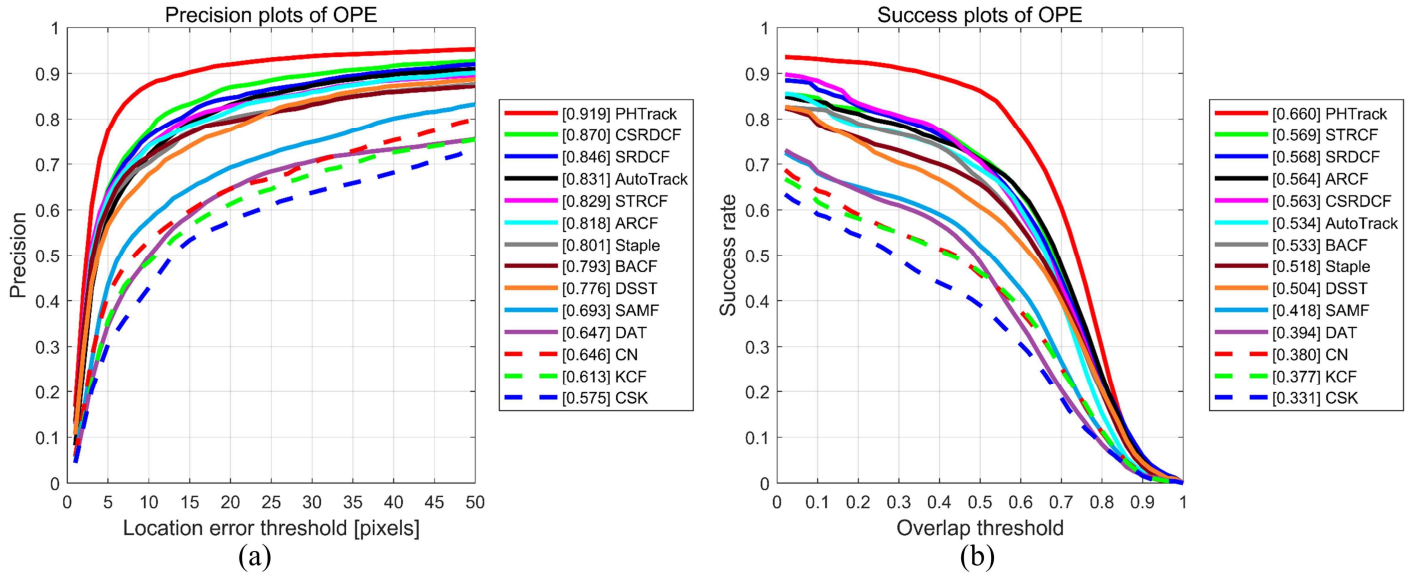


Fig. 6. Comparison with hand-crafted feature-based trackers on RGB videos. (a) Precision plot. (b) Success plot.

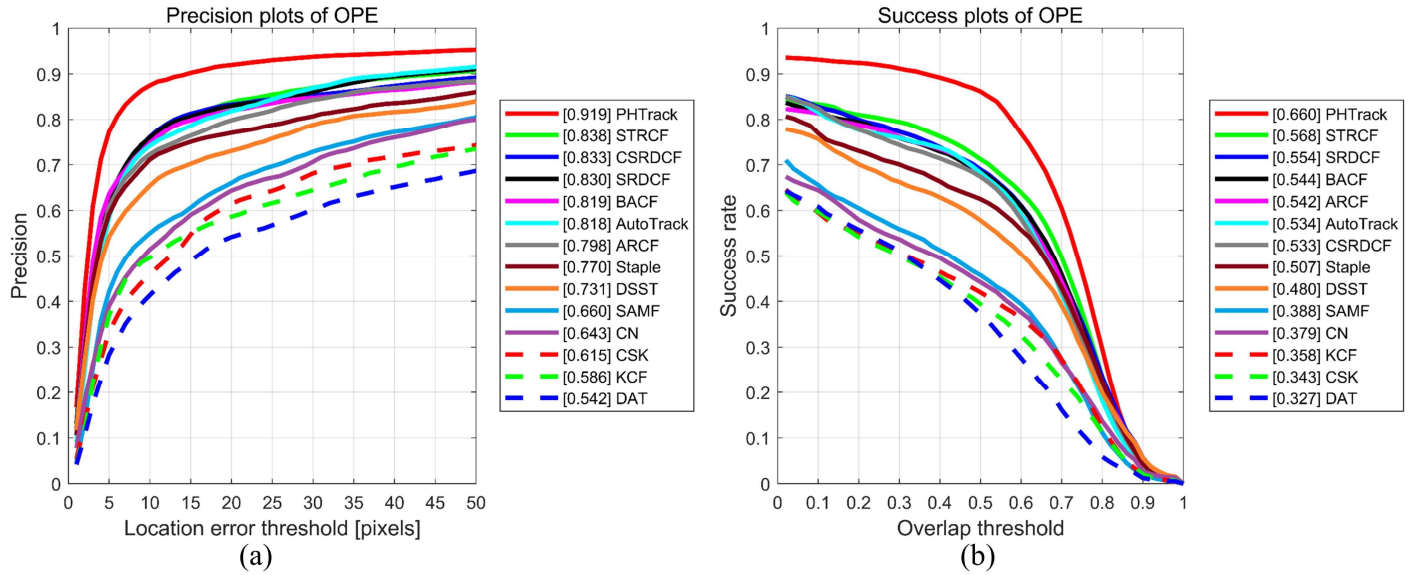


Fig. 7. Comparison with hand-crafted feature-based trackers on false color videos. (a) Precision plot. (b) Success plot.

bands), and RGB video (3 bands). The false color video is derived from the corresponding HS video, while the RGB video is captured from a viewpoint similar to the HS video. The HOTC dataset, characterized by 11 attributes, i.e., Occlusion (OCC), Scale Variation (SV), Fast Motion (FM), Motion Blur (MB), Illumination Variation (IV), Low Resolution (LR), Background Clutter (BC), In-Plane Rotation (IPR), Out-of-Plane Rotation (OPR), Out-of-View (OV), and Deformation (DEF), enables a synthetic evaluation of trackers.

2) Implementation Details: The proposed PHTrack is coded in Python with PyTorch 2.0.0 and trained on two NVIDIA RTX 3090 GPU cards. The foundation model is pre-trained on RGB datasets including ImageNet-VID [28], ImageNet-DET [28], YouTube-BB [51], and COCO [52]. Drawing inspiration from prompt learning, the prompt tuning is implemented on HS sets with a stochastic gradient descent optimizer involving an initial learning rate of 0.001, a batch size of 32, and a span of 20 epochs. The search and template patches are set to 255 and 127 pixels with 16 bands.

3) Assessment Metrics: As common metrics, the precision and success plots are utilized to benchmark trackers in one-pass evaluation [53]. The precision plot records the percentage of frames with a center location error v less than thresholds ranging from 1 to 50 pixels. Here, v is defined as $v = \sqrt{(x-X)^2 + (y-Y)^2}$, where (x, y) and (X, Y) denote the center of the estimated bounding box r_t and the ground truth r_g, respectively. In the success plot, the success rate aims to calculate the percentage of successful frames where the overlap score s surpasses thresholds ranging from 0 to 1. s is defined as $s = |r_t \cap r_g| / |r_t \cup r_g|$, where $\cup$ and $\cap$ are the union and intersection, and $|\cdot|$ stands for the number of pixels in a given region. As the guideline, trackers are ranked by the precision at 20 pixels on the precision plot and the area under the curve on the success plot, i.e., Pre and Suc, respectively. FPS is used to evaluate tracking speed.

TABLE I

PARALLEL COMPARISON WITH SOTA RGB TRACKERS

NO.	Tracker	Venue	Feature/Backbone/Tag	RGB		FAC/HS		PreD	SucD
				Pre	Suc	Pre	Suc		
1	CSK [54]	ECCV 2012	I	0.575	0.331	0.615	0.343	-4.0%	-1.2%
2	CN [55]	CVPR 2014	CN+I	0.646	0.380	0.643	0.379	0.3%	0.1%
3	SAMF [56]	ECCV 2015	HOG+CN+I	0.693	0.418	0.660	0.388	3.3%	**3.0%**
4	DAT [57]	CVPR 2015	CH	0.647	0.394	0.542	0.327	**10.5%**	**6.7%**
5	KCF [2]	TPAMI 2015	HOG	0.613	0.377	0.586	0.358	2.7%	1.9%
6	SRDCF [58]	ICCV 2015	HOG	**0.846**	**0.568**	**0.830**	**0.554**	1.6%	1.4%
7	Staple [59]	CVPR 2016	HOG+CN	0.801	0.518	0.770	0.507	3.1%	1.1%
8	DSST [60]	TPAMI 2017	HOG+I	0.776	0.504	0.731	0.480	**4.5%**	2.4%
9	BACF [61]	ICCV 2017	HOG	0.793	0.533	0.819	**0.544**	-2.6%	-1.1%
10	CSRDCF [62]	IJCV 2018	HOG+CN+CH	**0.870**	0.563	**0.833**	0.533	**3.7%**	**3.0%**
11	STRCF [63]	CVPR 2018	HOG+CN	0.829	**0.569**	**0.838**	**0.568**	-0.9%	0.1%
12	ARCF [64]	ICCV 2019	HOG+CN+I	0.818	**0.564**	0.798	0.542	2.0%	2.2%
13	AutoTrack [65]	CVPR 2020	HOG+CN+I	**0.831**	0.534	0.818	0.534	1.3%	0.0%
14	ECO [66]	CVPR 2017	VGG-M	0.872	0.577	0.834	0.556	3.8%	2.1%
15	SiamRPN [67]	CVPR 2018	AlexNet	0.902	0.592	0.757	0.486	**14.5%**	**10.6%**
16	DaSiamRPN [68]	ECCV 2018	AlexNet	0.878	0.622	0.850	0.575	2.8%	4.7%
17	ATOM [69]	CVPR 2019	ResNet-18	**0.917**	0.614	**0.867**	0.556	5.0%	5.8%
18	SiamRPN++ [70]	CVPR 2019	ResNet-50	**0.912**	**0.653**	0.847	**0.591**	6.5%	6.2%
19	UpdateNet [71]	ICCV 2019	AlexNet/DaSiamRPN	0.863	0.595	0.833	0.551	3.0%	4.4%
20	SiamDW [72]	CVPR 2019	CIRNext22/SiamFC	0.872	0.565	0.812	0.529	6.0%	3.6%
21	SiamMask [73]	CVPR 2019	ResNet-50	0.877	0.611	0.813	0.554	6.4%	5.7%
22	PrDiMP [74]	CVPR 2020	ResNet-50	**0.917**	0.634	0.829	0.565	8.8%	6.9%
23	SiamBAN [75]	CVPR 2020	ResNet-50	0.853	0.610	**0.863**	**0.587**	-1.0%	2.3%
24	SiamFC++ [76]	AAAI 2020	GoogLeNet	0.865	0.635	0.820	0.578	4.5%	5.7%
25	SiamGAT [77]	CVPR 2021	GoogLeNet	0.889	**0.649**	0.820	0.576	6.9%	**7.3%**
26	LightTrack [78]	CVPR 2021	Custom	0.814	0.593	0.761	0.530	5.3%	6.3%
27	Stark [79]	ICCV 2021	ResNet-50/ST	0.900	**0.637**	0.814	0.579	8.6%	5.8%
28	SiamCAR [3]	IJCV 2022	ResNet-50	0.882	0.636	0.846	**0.586**	3.6%	5.0%
29	OSTrack [80]	ECCV 2022	ViT-Base/256GOT	0.897	0.621	0.818	0.558	7.9%	6.3%
30	SBT [81]	CVPR 2022	SBT-Base/GOT	0.869	0.600	0.778	0.519	**9.1%**	**8.1%**
31	GRM [82]	CVPR 2023	ViT-Base/256GOT	**0.905**	0.630	0.815	0.566	**9.0%**	6.4%
32	SeqTrack [83]	CVPR 2023	ViT-Large/256GOT	0.890	0.612	**0.860**	0.583	3.0%	2.9%
33	SMAT [84]	WACV 2024	MobileViTv2	0.894	**0.637**	0.831	0.581	6.3%	5.6%
34	**PHTrack**	Ours	ResNet-50	n/a	n/a	**0.919**	**0.660**	n/a	n/a

The top three scores are marked in red, green, and blue. RGB, FAC, and HS are the red-green-blue, false color, and hyperspectral videos, respectively. PreD and SucD are the Pre degradation and Suc degradation from RGB to false color videos. Trackers using hand-crafted features are shown above the dashed line, while trackers using deep features are presented below. n/a stands for not applicable.

B. Comparison with SOTA RGB Trackers

In this section, we conduct a comparative analysis of PHTrack with 33 SOTA RGB trackers. To facilitate clarity, we categorize these trackers into two groups: (i) those based on hand-crafted features and (ii) those based on deep features. These trackers encompass a wide range of features, backbones, and paradigms.

1) Hand-crafted Feature-based Trackers: We compare PHTrack with 13 RGB SOTAs based on hand-crafted features, including CSK [54], CN [55], SAMF [56], DAT [57], KCF [2], SRDCF [58], Staple [59], DSST [60], BACF [61], CSRDCF [62], STRCF [63], ARCF [64], and AutoTrack [65]. These trackers are assessed on RGB and false color videos, while PHTrack is

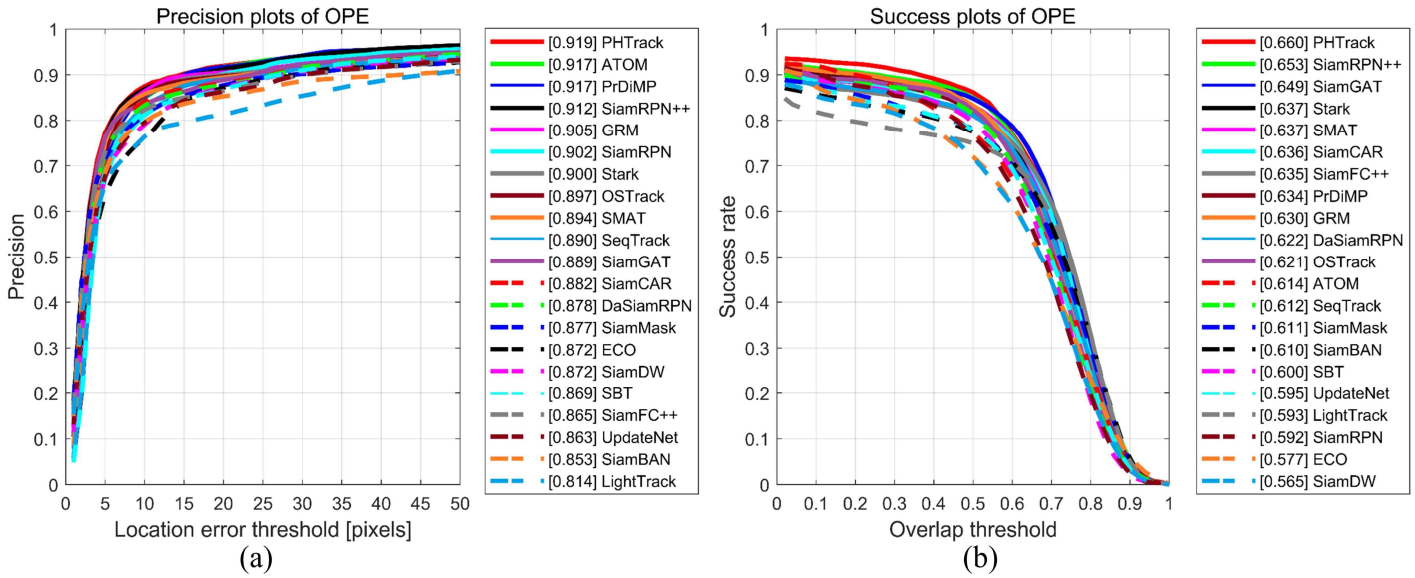


Fig. 8. Comparison with deep feature-based trackers on RGB videos. (a) Precision plot. (b) Success plot.

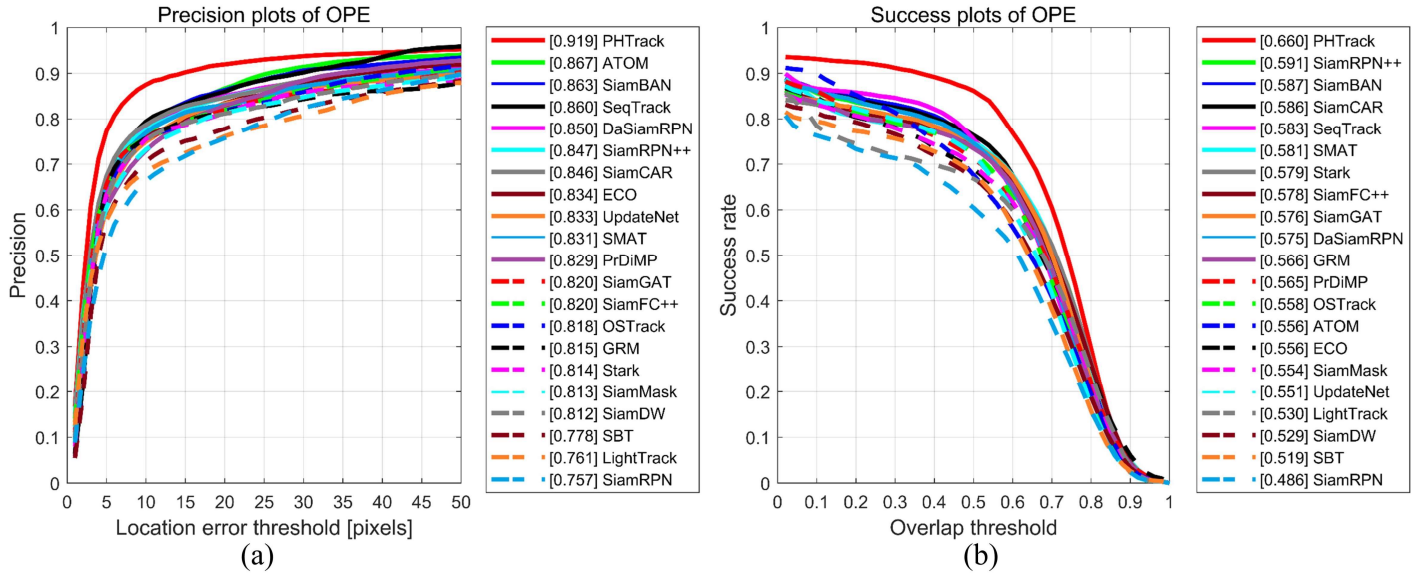


Fig. 9. Comparison with deep feature-based trackers on false color videos. (a) Precision plot. (b) Success plot.

tested on HS videos. Table I details the characteristics and results. Fig. 6 presents the precision and success plots tested on RGB videos.

PHTrack yields optimal results with Pre of 0.919 and Suc of 0.660. Compared to CSRDCF, PHTrack exhibits gains of 4.9% in Pre and 9.7% in Suc. Compared to SRDCF and STRCF, PHTrack achieves impressive improvements in Suc of 9.2% and 9.1%, respectively. The finding underlines the potential of leveraging abundant material information present in HS data.

It is conceivable to employ the RGB tracker by converting the HS video into a false color video. Naturally, we conduct experiments on the false color video, as illustrated in Fig. 7 and Table I. STRCF maintains a respectable performance, followed by SRDCF and BACF with Suc scores of 0.568, 0.554, and 0.544, respectively. Remarkably, PHTrack emerges as the frontrunner, outperforming its counterparts by 9.2%, 10.6%, and 11.6%, respectively. This is attributed to the MOP and DIP, which stimulate foundation models and distill cross-modal features.

2) Deep Feature-based Trackers: Deep features, known for their discriminative capabilities, have made significant strides in the field of RGB tracking. Here we evaluate PHTrack against 20 deep feature-based trackers including ECO [66], SiamRPN [67], DaSiamRPN [68], ATOM [69], SiamRPN++ [70], UpdateNet [71], SiamDW [72], SiamMask [73], PrDiMP [74], SiamBAN [75], SiamFC++ [76], SiamGAT [77], LightTrack [78], Stark [79], SiamCAR [3], OSTrack [80], SBT [81], GRM [82], SeqTrack [83], and SMAT [84].

Detailed results are presented in Table I, while precision and success plots, tested on RGB and false color videos, are shown in Fig. 8 and Fig. 9, respectively. Overall, leveraging efficient and effective prompter, i.e., MOP and DIP, PHTrack demonstrates impressive results.

TABLE II

CHARACTERISTICS AND RESULTS OF HS TRACKERS

NO.	Tracker	Venue	Framework	Feature	UFB	MOP	FPS	Pre	Suc
1	CNHT [35]	ICSM 2018	KCF	Deep feature	✓	CPU	2.6	0.336	0.171
2	DeepHKCF [87]	TGRS 2019	KCF	Deep feature	-	CPU	0.9	0.543	0.303
3	MHT [18]	TIP 2020	KCF	Hand-crafted feature	✓	CPU	2.2	0.883	0.586
4	BAE-Net [88]	ICIP 2020	VITAL	Deep feature	✓	GPU	0.5	0.879	0.606
5	MFI [36]	WISP 2021	KCF	Deep + Hand-crafted features	-	CPU	0.4	**0.893**	0.601
6	SST-Net [19]	WISP 2021	VITAL	Deep feature	✓	GPU	0.5	**0.917**	0.623
7	TSCFW [17]	TGRS 2022	KCF	Hand-crafted features	✓	CPU	3.4	0.887	0.604
8	TASSCF [16]	CVIU 2022	KCF	Hand-crafted features	-	CPU	**16.0**	0.870	0.587
9	DeepTASSCF [16]	CVIU 2022	KCF	Deep feature	-	CPU	6.0	0.875	0.602
10	SiamOHOT [22]	TGRS 2023	SiamFC	Deep feature	✓	GPU	**38.0**	0.884	**0.634**
11	SiamBAG [20]	TGRS 2023	SiamFC	Deep feature	✓	GPU	5.7	**0.893**	**0.632**
12	SiamHT [30]	NCA 2023	SiamFC	Deep feature	-	GPU	**16.0**	0.878	0.620
13	**PHTrack**	Ours	SiamFC	Deep feature	✓	GPU	**15.2**	**0.919**	**0.660**

UFB denotes the attempt to use of full band. MOP denotes the main operation platform. WISP denotes the WHISPERS.

C. Parallel Analysis with RGB Trackers

Fig. 10 compares the top 13 trackers using hand-crafted and deep features across RGB and false color videos, providing critical insights that advance understanding in HS tracking.

(i) Deep feature-based trackers outperform hand-crafted ones by learning complex representations from training data. This shift reduces the reliance on expensive hand-crafted feature design, driving progress in HS object tracking research.

(ii) Performance on false-color videos is lower than on RGB videos, though rankings remain consistent.

(iii) Deep feature-based trackers show greater performance degradation (PreD and SucD) compared to hand-crafted feature-based counterparts due to their heavy reliance on specific characteristics present in RGB format training data.

(iv) Converting HS video to the false color format leads to inevitable loss and distortion of crucial material information, hindering robust performance.

(v) Leveraging physical material cues in the HS format is expected to improve tracking performance. Additionally, video super-resolution techniques [85], [86] may further enhance HS object tracking.

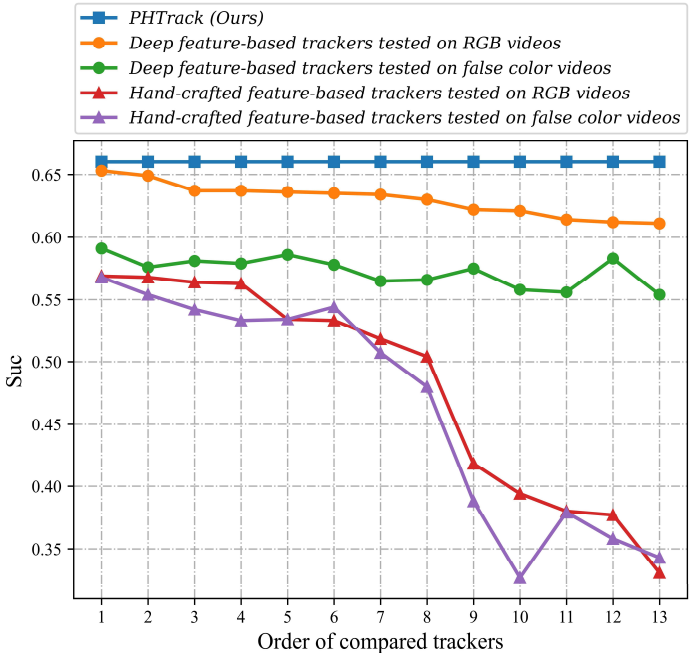


Fig. 10. Parallel comparisons with RGB trackers utilizing hand-crafted and deep features, ranked by Suc on RGB videos. Hand-crafted trackers: STRCF, SRDCF, ARCF, CSRDCF, AutoTrack, BACF, Staple, DSST, SAMF, DAT, CN, KCF, CSK. Deep feature-based trackers: SiamRPN++, SiamGAT, SMAT, Stark, SiamCAR, SiamFC++, PrDiMP, GRM, DaSiamRPN, OSTrack, ATOM, SeqTrack, SiamMask.

D. Comparison with HS Trackers

Furthermore, we compare PHTrack with 12 SOTA HS trackers including CNHT [35], DeepHKCF [87], MHT [18], BAE-Net [88], MFI [36], SST-Net [19], TSCFW [17], TASSCF [16], DeepTASSCF [16], SiamOHOT [22], SiamBAG [20], and SiamHT [30]. Their characteristics and experimental results are shown in Table II, while the precision and success plots are depicted in Fig. 11. Regarding Pre, PHTrack, SST-Net, MFI, and SiamBAG achieve top rankings with scores of 0.919, 0.917, 0.893, and 0.893, respectively. In terms of Suc, SiamOHOT, SiamBAG, SST-Net, and SiamHT demonstrate competitive outcomes with scores of 0.634, 0.632, 0.623, and 0.620 respectively, securing the top four positions among the compared trackers. PHTrack surpasses them by 2.6%, 2.8%, 3.7%, and 4.0%, respectively. Overall, PHTrack showcases optimal performance in both Pre and Suc, underscoring the potential of prompt learning in HS video object tracking. Furthermore, Table II highlights that leading trackers (e.g., SiamOHOT, SiamBAG, SST-Net, and PHTrack) typically inherit the Siamese network, whereas low-ranked trackers (e.g., DeepTASSCF and CNHT) rely on the kernelized correlation filter. This suggests that the

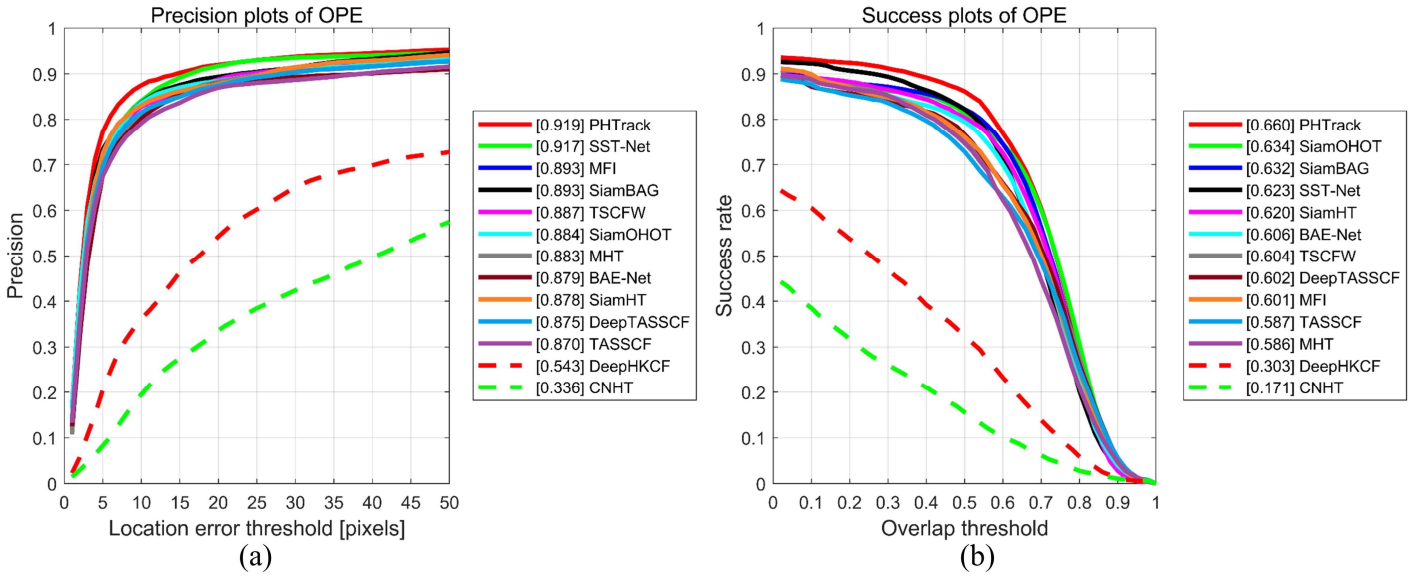


Fig. 11. Comparison with HS trackers on HS videos. (a) Precision plot. (b) Success plot.

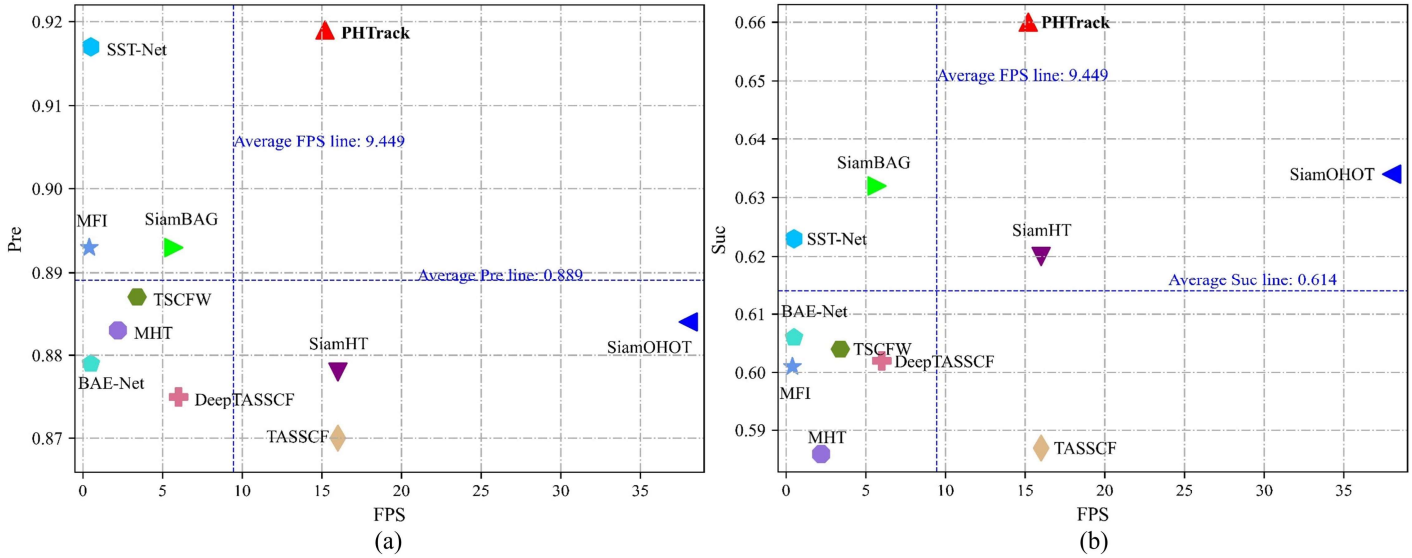


Fig. 12. Accuracy-speed comparison with HS trackers on HS videos. (a) Pre vs. FPS. (b) Suc vs. FPS.

Siamese network may outperform kernelized correlation filter in HS video tracking, similar to trends in RGB tracker evolution (see Table I). Recent studies, such as SiamOHOT, SiamBAG, and SST-Net, harness rich spectral cues for effective representations, as abundant spectral cues enhance the tracker's material identification, leading to improved performance. The PHTrack, inheriting the Siamese network and integrating learned prompters, achieves significant improvements.

E. Accuracy VS. Speed on HS Videos

Fig. 12 illustrates the trade-off between accuracy (Pre and Suc) and speed (FPS) for various HS trackers, including MHT [18], BAE-Net [88], MFI [36], SST-Net [19], TSCFW [17], TASSCF [16], DeepTASSCF [16], SiamOHOT [22], SiamBAG [20], SiamHT [20], and PHTrack. In Fig. 12(a), most trackers have either accuracy below the average Pre of 0.889 (e.g., MHT, SiamHT, DeepTASSCF) or speed below the average FPS of 9.449 (e.g., SiamBAG, TSCFW, MFI). Only PHTrack exceeds both average Pre and FPS. Specifically, SiamOHOT attains the highest speed, followed by SiamHT, TASSCF, and PHTrack. However, SiamOHOT's Pre is 0.884, 3.5% lower than PHTrack. In Fig. 12(b), PHTrack, SiamHT, and SiamOHOT surpass the average Suc of 0.614 and FPS. PHTrack has the highest Suc at 0.660, outperforming SiamHT and SiamOHOT by 4.0% and 2.6% respectively. SiamOHOT emphasizes knowledge distillation for enhanced tracking efficiency. SEE-Net [26] employs decision-level fusion to improve effectiveness at the cost of additional computational burden thus leading to low speed (8.72 FPS). PHTrack inherits the prompt learning paradigm and feature-level fusion strategy with the introduction of a moderate amount of learnable parameters to present a favorable accuracy-speed trade-off, positioning it as an effective candidate for HS video tracking.

F. Attribute-based Evaluations

We perform attribute-based evaluations with 11 RGB trackers, i.e., BACF [61], CSRDCF [62], ARCF [64], AutoTrack [65],

TABLE III

PRE RESULTS FOR EACH ATTRIBUTE AND OVERALL

Tracker	Venue	BC	DEF	FM	IPR	IV	LR	MB	OCC	OPR	OV	SV	OVE
BACF [61]	ICCV 2017	0.762	**0.947**	0.774	0.867	0.804	0.777	0.855	0.779	0.887	0.900	0.838	0.819
CSRDCF [62]	IJCV 2018	0.791	**0.956**	0.758	0.856	0.783	0.820	0.820	0.796	0.865	0.878	0.845	0.833
ARCF [64]	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
AutoTrack [65]	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
LightTrack [78]	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
Stark [79]	ICCV 2021	0.791	0.917	0.793	0.921	0.848	0.706	0.749	0.813	0.932	**1.000**	0.852	0.814
SiamCAR [3]	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
OSTrack [80]	ECCV 2022	0.719	0.909	0.949	0.826	0.841	0.831	0.929	0.812	0.911	**1.000**	0.869	0.818
SBT [81]	CVPR 2022	0.659	0.878	0.930	0.746	0.840	0.806	0.939	0.773	0.840	**0.959**	0.821	0.778
GRM [82]	CVPR 2023	0.743	0.905	**0.955**	0.832	0.789	0.819	**0.947**	0.785	0.944	**1.000**	0.876	0.815
SMAT [84]	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
CNHT [35]	ICSM 2018	0.313	0.448	0.403	0.487	0.251	0.127	0.256	0.218	0.419	0.412	0.287	0.336
DeepHKCF [87]	TGRS 2019	0.513	0.636	0.500	0.737	0.304	0.304	0.626	0.439	0.629	0.548	0.501	0.543
MHT [18]	TIP 2020	0.901	0.908	0.774	0.940	0.806	0.827	0.839	0.816	0.893	0.887	0.874	0.883
MFI [36]	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**
TSCFW [17]	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**
TASSCF [16]	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
DeepTASSCF [16]	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
SiamOHOT [22]	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
SiamBAG [20]	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**
SiamHT [30]	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
PHTrack	Ours	**0.925**	0.919	**0.993**	**0.952**	0.868	**0.894**	**0.986**	**0.882**	0.947	0.887	**0.916**	**0.919**

TABLE IV

SUC RESULTS FOR EACH ATTRIBUTE AND OVERALL

Tracker	Venue	BC	DEF	FM	IPR	IV	LR	MB	OCC	OPR	OV	SV	OVE
BACF [61]	ICCV 2017	0.519	0.672	0.568	0.631	0.464	0.417	0.596	0.523	0.643	0.525	0.544	0.544
CSRDCF [62]	IJCV 2018	0.541	0.662	0.577	0.587	0.419	0.446	0.585	0.509	0.593	0.330	0.513	0.533
ARCF [64]	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
AutoTrack [65]	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
LightTrack [78]	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
Stark [79]	ICCV 2021	0.566	0.695	0.539	0.696	**0.570**	0.483	0.538	0.575	0.704	**0.783**	0.605	0.579
SiamCAR [3]	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
OSTrack [80]	ECCV 2022	0.486	0.668	0.613	0.607	0.550	0.513	0.613	0.565	0.664	**0.779**	0.594	0.558
SBT [81]	CVPR 2022	0.429	0.641	0.583	0.528	0.547	0.507	0.624	0.524	0.604	0.728	0.559	0.519
GRM [82]	CVPR 2023	0.516	0.675	0.613	0.625	0.528	0.517	0.630	0.560	0.702	**0.810**	0.612	0.566
SMAT [84]	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
CNHT [35]	ICSM 2018	0.183	0.288	0.176	0.272	0.097	0.027	0.094	0.118	0.259	0.144	0.156	0.171
DeepHKCF [87]	TGRS 2019	0.284	0.426	0.264	0.485	0.129	0.083	0.363	0.250	0.428	0.286	0.298	0.303
MHT [18]	TIP 2020	0.606	0.664	0.542	0.670	0.477	0.475	0.560	0.564	0.644	0.626	0.574	0.586
MFI [36]	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
TSCFW [17]	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
TASSCF [16]	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
DeepTASSCF [16]	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
SiamOHOT [22]	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**
SiamBAG [20]	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**
SiamHT [30]	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
PHTrack	Ours	**0.679**	0.674	**0.735**	**0.725**	**0.578**	**0.583**	**0.730**	**0.618**	0.709	0.604	**0.646**	**0.660**

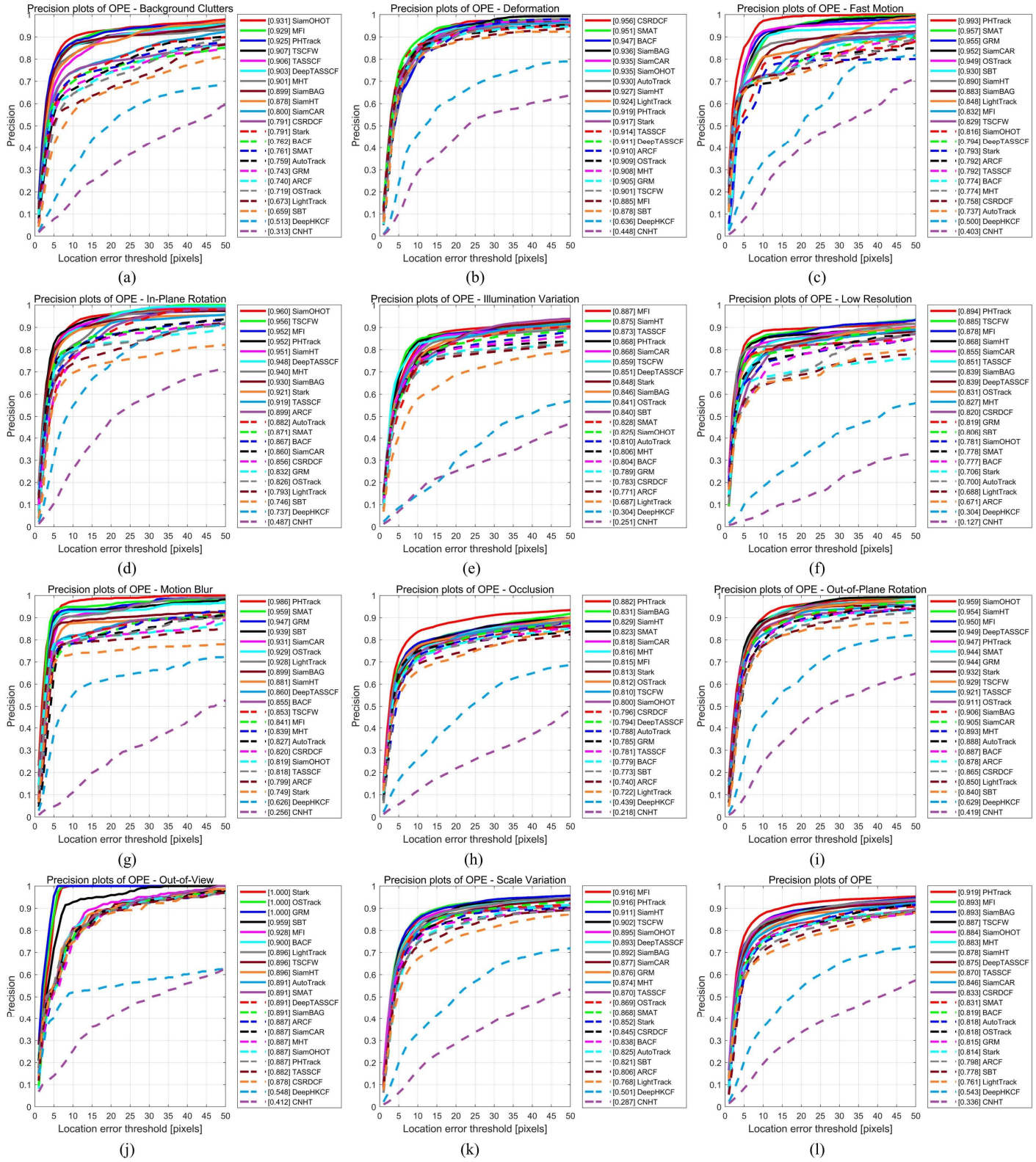


Fig. 13. Precision plots for each attribute and overall. (a) BC. (b) DEF. (c) FM. (d) IPR. (e) IV. (f) LR. (g) MB. (h) OCC. (i) OPR. (j) OV. (k) SV. (l) OVE.

LightTrack [78], Stark [79], SiamCAR [3], OSTrack [80], SBT [81], GRM [82], and SMAT [84] along with 10 SOTA HS trackers including CNHT [35], DeepHKCF [87], MHT [18], MFI [36], TSCFW [17], TASSCF [16], DeepTASSCF [16], SiamOHOT [22], SiamBAG [20], and SiamHT [30]. The evaluation tests RGB trackers on false color videos derived from HS videos. Tables III and IV showcase the Pre and Suc scores for each attribute and overall (OVE). Additionally, Fig. 13 and Fig. 14 display the precision plot and success plot. PHTrack ranks among the top three for seven attributes (BC, FM, IPR, LR, MB, OCC, and SV) out of the 11 evaluated attributes and yields the first overall for Pre. In terms of Suc, PHTrack ranks in the top

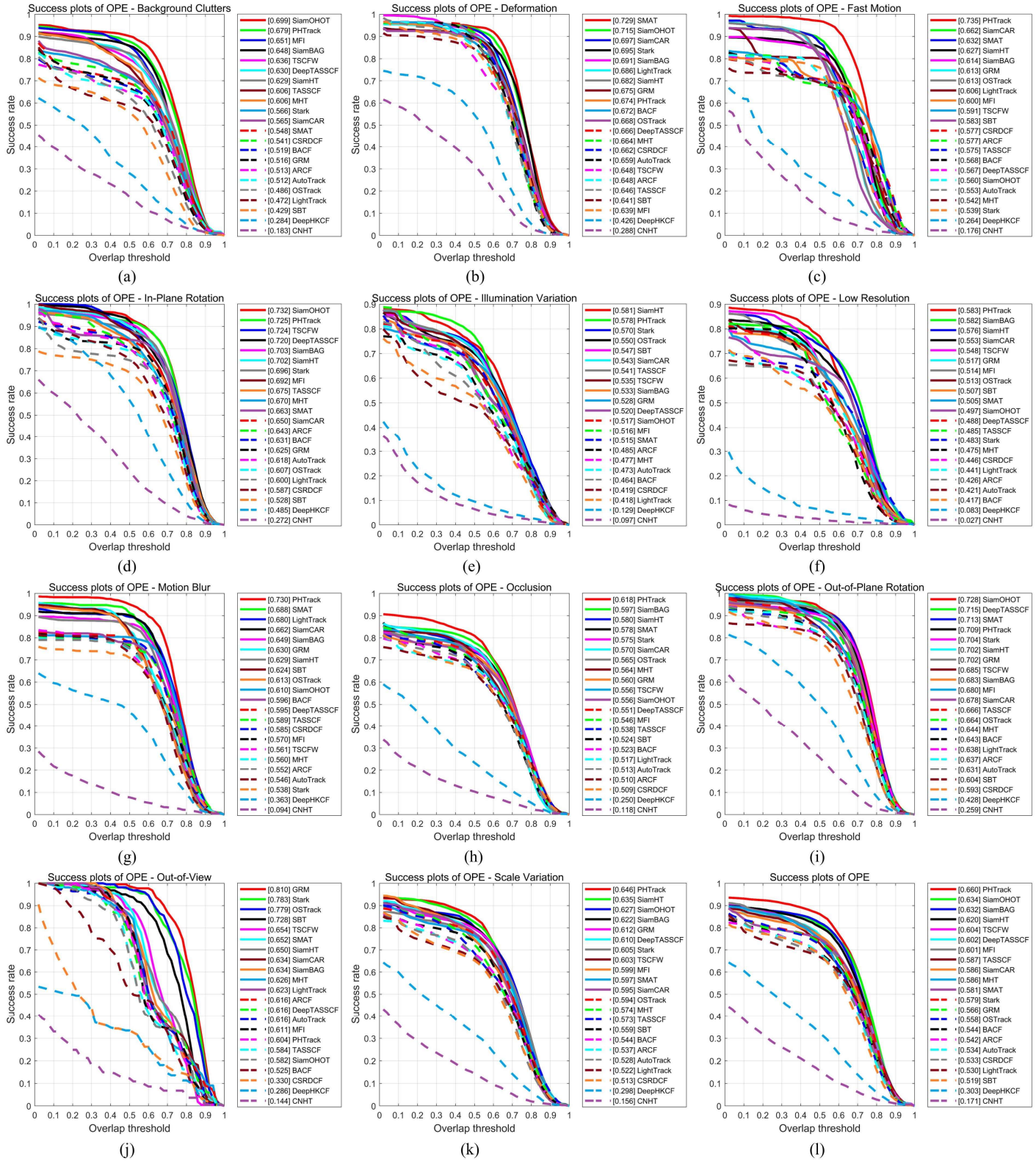


Fig. 14. Success plots for each attribute and overall. (a) BC. (b) DEF. (c) FM. (d) IPR. (e) IV. (f) LR. (g) MB. (h) OCC. (i) OPR. (j) OV. (k) SV. (l) OVE.

two for eight attributes (BC, FM, IPR, IV, LR, MB, OCC, and SV) and achieves first place overall. LR poses a significant challenge. Benefiting from MOP and DIP, PHTrack effectively uses physical material cues to tackle this attribute, achieving remarkable Pre and Suc scores of 0.894 and 0.583, respectively. OCC presents another tough attribute where objects are partially or fully occluded. Notably, PHTrack attains the highest Suc of 0.618, outperforming SiamHT and SiamBAG by 3.8% and 2.1%, respectively, owing to its robust discrimination capabilities empowered by MOP and DIP. In conclusion, extensive experiments validate PHTrack's efficacy in challenging scenes and highlight the role of prompt learning in HS tracking.

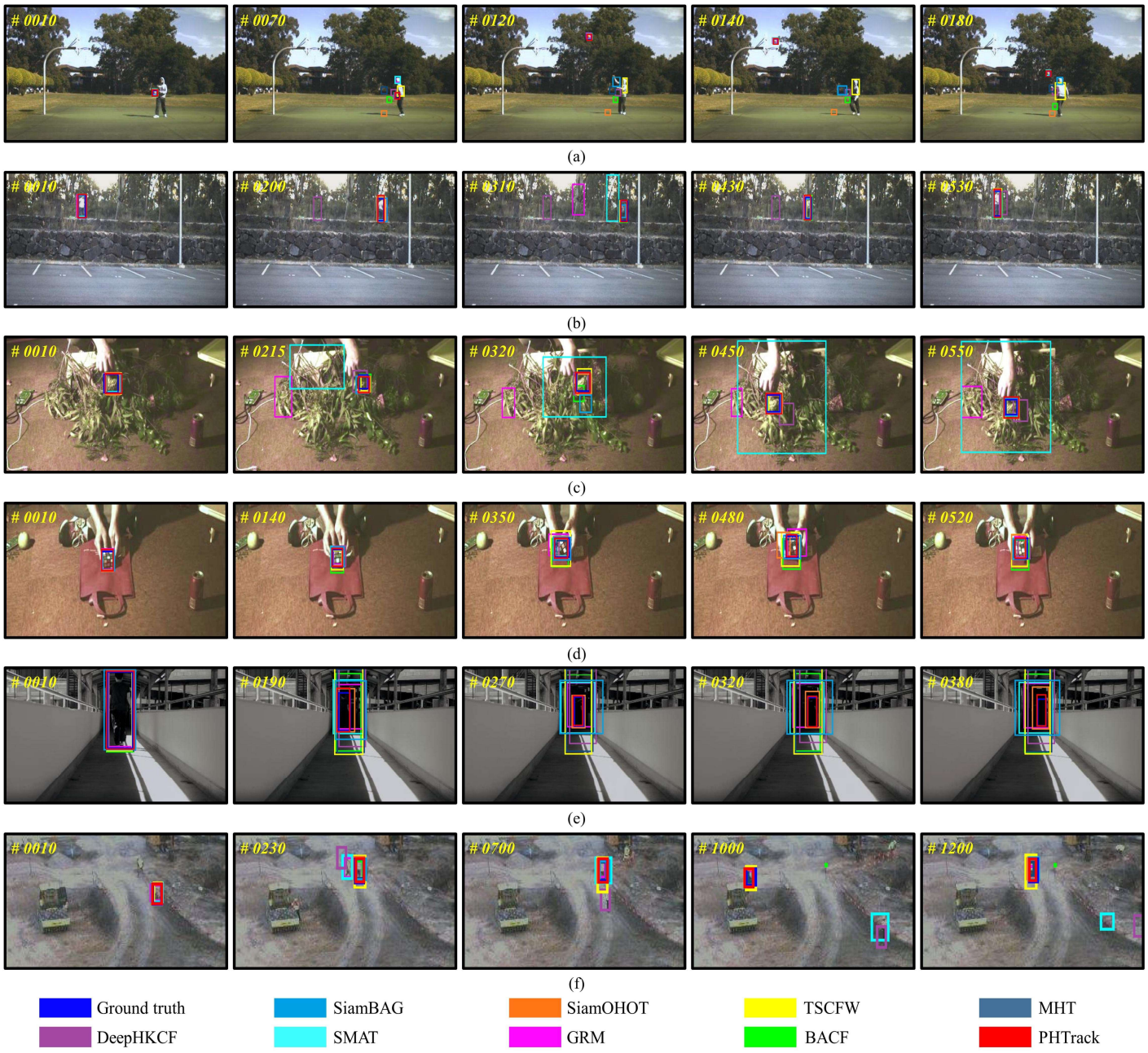


Fig. 15. Qualitative results. (a) *basketball*, attribute: FM, MB, OCC, and LR. (b) *forest*, attribute: BC and OCC. (c) *fruit*, attribute: BC and OCC. (d) *rubik*, attribute: DEF, IPR, and OPR. (e) *student*, attribute: IV and SV. (f) *worker*, attribute: SV, BC, and LR. Bounding boxes are displayed in the false color image. The current frame is shown in the top-left corner.

G. Visual Comparisons

Fig. 15 displays visual samples of trackers including BACF [61], GRM [82], SMAT [84], DeepHKCF [87], MHT [18], TSCFW [17], SiamOHOT [22], SiamBAG [20], and PHTrack. Benefiting from MOP and DIP, PHTrack excels at mitigating interferences and precisely determining object states such as position and scale. Qualitative findings illustrate that PHTrack consistently delivers high accuracy and robustness across diverse scenarios, making it an ideal candidate for HS object tracking applications.

H. Ablation Studies

The key contributions of PHTrack encompass MOP and DIP. To validate their impact, we evaluate five models: Model-1, Model-2, Model-3, Model-4, and Model-5. Notably, the baseline Model-1 is performed on HS videos by transforming them into false color videos, whereas the remaining models are tested on HS videos.

1) Effectiveness of MOP: We conduct comparisons to evaluate the effectiveness of incorporating the multi-scale (MS) and WST architectures within the MOP. Detailed results and compositions are presented in Table V. Using Model-1 as a baseline, Model-2 and Model-3 introduce MS and WST, respectively. With the addition of MS, Model-2 shows significant improvements in Pre and Suc by 6.6% and 7.1%, respectively, over Model-1. Similarly, incorporating MS in Model-3 shows gains of 1.6% in Pre and 0.9% in Suc, as shown in PHTrack. Furthermore, the addition of WST in Model-3 results in notable improvements of

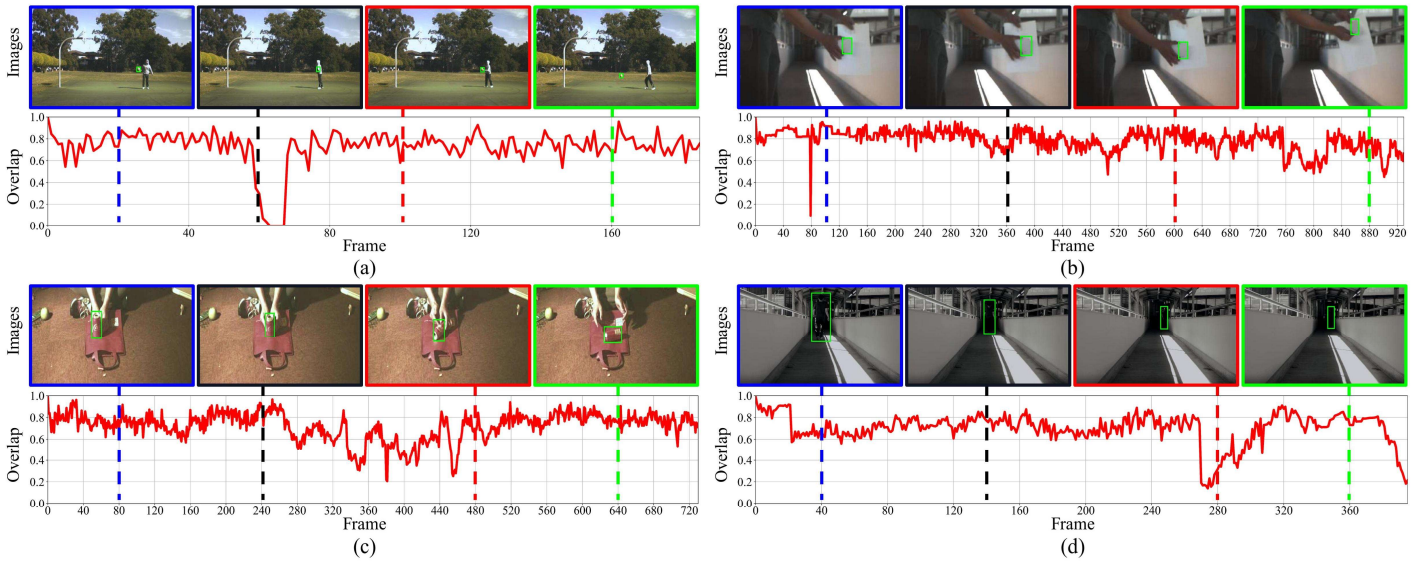


Fig. 16. Overlap curves and tracking results of PHTrack. (a) *basketball*, attribute: OCC, FM, MB, and LR. (b) *card*, attribute: OCC and SV. (c) *coke*, attribute: BC, IPR, OPR, FM, and SV. (d) *student*, attribute: SV and IV. Tracked results are marked in green.

TABLE V

ABLATION EXPERIMENT ON THE EFFECTIVENESS OF MOP

Model	MOP		Pre	Suc	PreI	SucI
	MS	WST				
Model-1	-	-	0.846	0.586	n/a	n/a
Model-2	✓	-	0.912	0.657	6.6%	7.1%
Model-3	-	✓	0.903	0.651	5.7%	6.5%
PHTrack	✓	✓	**0.919**	**0.660**	**7.3%**	**7.4%**

PreI and SucI are the Pre and Suc improvements of the current model compared to Model-1.

TABLE VI

ABLATION EXPERIMENT ON THE EFFECTIVENESS OF DIP

Model	SUM	CAT	DIP	Pre	Suc	PreI	SucI
Model-1	-	-	-	0.846	0.586	n/a	n/a
Model-4	✓	-	-	0.801	0.545	-4.5%	-4.1%
Model-5	-	✓	-	0.803	0.537	-4.3%	-4.9%
PHTrack	-	-	✓	**0.919**	**0.660**	**7.3%**	**7.4%**

5.7% and 6.5% in Pre and Suc, respectively, when compared to Model-1. Conversely, omitting WST in Model-2 leads to a decrease of 0.7% in Pre and 0.3% in Suc compared to PHTrack. A comparison between PHTrack and Model-1 demonstrates remarkable boosts of 7.3% in Pre and 7.4% in Suc. Experimental results underscore the efficacy of MOP, enabling PHTrack to exploit the foundation model and prior knowledge, consistently improving performance.

2) Effectiveness of DIP: This section proves the effectiveness of DIP, as shown in Table VI. Common fusion strategies, i.e., element-wise summation (SUM) and concatenation (CAT), are incorporated into Model-4 and Model-5 for comparison with PHTrack with DIP. A comparison between PHTrack and Model-4 reveals enhancements of 11.8% and 11.5% in Pre and Suc when SUM is replaced by DIP. Comparing Model-5 and PHTrack reveals a decrease of 11.6% and 12.3% in Pre and Suc when CAT is used instead of DIP. The challenges with element-wise summation and concatenation lie in their inability to effectively handle the significance of multi-modal features across various spatial and channel contexts. In contrast, DIP excels in treating features with diverse spatial and channel locations, thereby stimulating the foundation model to achieve competitive outcomes. Fig. 16 depicts overlap curves and tracking results of PHTrack, showcasing its capacity to sustain high overlap scores despite challenging scenes. Extensive experiments have revealed the beneficial effects of MOP and DIP, supporting the competitive performance demonstrated by PHTrack.

V. CONCLUSION

In this study, we design a Prompting for Hyperspectral Video Tracking framework, denoted as PHTrack. The goal is to alleviate data anxiety by efficiently leveraging the capabilities of the foundation model through prompt learning. Specifically, an MOP is proposed to learn multi-modal generation from HS images to extract rich spectral information. Guided by MOP, our framework bridges band gaps, enhances model adaptation, and leverages prior knowledge. Additionally, a DIP is developed to integrate cross-modal features, facilitating feature fusion and efficiently addressing the challenge of huge volumes. Extensive experiments confirm the effectiveness of the proposed method.

Nevertheless, the computational demands inherent in the current framework represent an obstacle in dynamic environments requiring instantaneous response times. This underscores the imperative for future research to focus on refining the model structure to achieve higher efficiency while maintaining accuracy. Furthermore, developing a unified spectral-spatial-temporal framework for HS tracking may be a promising avenue.

REFERENCES

[1] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep Learning for Visual Tracking: A Comprehensive Survey," *IEEE Trans. Intell. Transp. Syst.,* vol. 23, no. 5, pp. 3943-3968, May 2022.

[2] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-Speed Tracking with Kernelized Correlation Filters," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 37, no. 3, pp. 583-96, Mar 2015.

[3] Y. Cui *et al.*, "Joint Classification and Regression for Visual Tracking with Fully Convolutional Siamese Networks," *Int. J. Comput. Vis.,* vol. 130, no. 2, pp. 550-566, Feb 2022.

[4] Y. Cui, C. Jiang, G. Wu, and L. Wang, "MixFormer: End-to-End Tracking With Iterative Mixed Attention," *IEEE Trans. Pattern Anal. Mach. Intell.,* pp. 1-18, 2024.

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

[6] Y. Cai, X. Sui, and G. Gu, "Multi-modal multi-task feature fusion for RGBT tracking," *Inf. Fusion,* Article vol. 97, Sep 2023, Art no. 101816.

[7] Xuefeng Zhu, Tianyang Xu, Zongtao Liu, Zhangyong Tang, Xiaojun Wu, and J. Kittler, "UniMod1K: Towards a More Universal Large-Scale Dataset and Benchmark for Multi-modal Learning," *Int. J. Comput. Vis.,* pp. 1-16, 02/22 2024.

[8] X. Wang *et al.*, "VisEvent: Reliable Object Tracking via Collaboration of Frame and Event Flows," *IEEE Trans. Cybern.,* vol. 54, no. 3, pp. 1997-2010, 2024.

[9] C. L. Li *et al.*, "LasHeR: A Large-Scale High-Diversity Benchmark for RGBT Tracking," *IEEE Trans. Image Process.,* vol. 31, pp. 392-404, 2022.

[10] X. Cheng *et al.*, "Deep Feature Aggregation Network for Hyperspectral Anomaly Detection," *IEEE Trans. Instrum. Meas.,* pp. 1-1, 2024.

[11] S. Lin, X. Cheng, Y. Zeng, Y. Huo, M. Zhang, and H. Wang, "Low-Rank and Sparse Representation Inspired Interpretable Network for Hyperspectral Anomaly Detection," *IEEE Trans. Instrum. Meas.,* pp. 1-1, 2024.

[12] Y. Xu, Y. Xu, H. Jiao, Z. Gao, and L. Zhang, "S³ANet: Spatial–Spectral Self-Attention Learning Network for Defending Against Adversarial Attacks in Hyperspectral Image Classification," *IEEE Trans. Geosci. Remote Sens.,* vol. 62, pp. 1-13, 2024.

[13] J. Li, K. Zheng, L. Gao, L. Ni, M. Huang, and J. Chanussot, "Model-Informed Multistage Unsupervised Network for Hyperspectral Image Super-Resolution," *IEEE Trans. Geosci. Remote Sens.,* vol. 62, pp. 1-17, 2024.

[14] T. Han, Y. Tang, Y. Chen, X. Yang, Y. Guo, and S. Jiang, "SDC-GAE: Structural Difference Compensation Graph Autoencoder for Unsupervised Multimodal Change Detection," *IEEE Trans. Geosci. Remote Sens.,* vol. 62, pp. 1-16, 2024.

[15] Y. Z. Chen, Q. Q. Yuan, Y. Q. Tang, Y. Xiao, J. He, and L. P. Zhang, "SPIRIT: Spectral Awareness Interaction Network With Dynamic Template for Hyperspectral Object Tracking," *IEEE Trans. Geosci. Remote Sens.,* vol. 62, 2024, Art no. 5503116.

[16] Y. Tang, Y. Liu, and H. Huang, "Target-aware and spatial-spectral discriminant feature joint correlation filters for hyperspectral video object tracking," *Comput. Vis. Image Underst.,* Article vol. 223, Oct 2022, Art no. 103535.

[17] Z. Hou, W. Li, J. Zhou, and R. Tao, "Spatial-Spectral Weighted and Regularized Tensor Sparse Correlation Filter for Object Tracking in Hyperspectral Videos," *IEEE Trans. Geosci. Remote Sens.,* Article vol. 60, 2022, Art no. 5541012.

[18] F. Xiong, J. Zhou, and Y. Qian, "Material Based Object Tracking in Hyperspectral Videos," *IEEE Trans. Image Process.,* vol. 29, pp. 3719-3733, Jan 15 2020.

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

[20] W. Li, Z. F. Hou, J. Zhou, and R. Tao, "SiamBAG: Band Attention Grouping-Based Siamese Object Tracking Network for Hyperspectral Videos," *IEEE Trans. Geosci. Remote Sens.,* vol. 61, 2023, Art no. 5514712.

[21] Y. Chen, Q. Yuan, Y. Tang, Y. Xiao, J. He, and Z. Liu, "SENSE: Hyperspectral Video Object Tracker via Fusing Material and Motion Cues," *Inf. Fusion,* vol. 109, p. 102395, 2024.

[22] C. Sun, X. Wang, Z. Liu, Y. Wan, L. Zhang, and Y. Zhong, "SiamOHOT: A Lightweight Dual Siamese Network for Onboard Hyperspectral Object Tracking via Joint Spatial-Spectral Knowledge Distillation," *IEEE Trans. Geosci. Remote Sens.,* vol. 61, pp. 1-12, 2023, Art no. 5521112.

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

[24] Y. Chen, Y. Tang, Q. Yuan, and L. Zhang, "REPS: Rotation equivariant Siamese network enhanced by probability segmentation for satellite video tracking," *Int. J. Appl. Earth. Obs. Geoinf.,* vol. 128, p. 103741, 2024.

[25] Y. Chen *et al.*, "Satellite video single object tracking: A systematic review and an oriented object tracking benchmark," *ISPRS J. Photogramm. Remote Sens.,* vol. 210, pp. 212-240, 2024.

[26] Z. Li, F. Xiong, J. Zhou, J. Lu, and Y. Qian, "Learning a Deep Ensemble Network With Band Importance for Hyperspectral Object Tracking," *IEEE Trans. Image Process.,* vol. 32, pp. 2901-2914, 2023.

[27] M. Muller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild," in *15th European Conference on Computer Vision (ECCV)*, Munich, GERMANY, vol. 11205, 2018, pp. 310-327.

[28] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.,* vol. 115, no. 3, pp. 211-252, Dec 2015.

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

[30] Y. Tang, H. Huang, Y. Liu, and Y. Li, "A Siamese network-based tracking framework for hyperspectral video," *Neural Comput. Appl.,* Article vol. 35, no. 3, pp. 2381-2397, Jan 2023.

[31] Y. Tang, Y. Liu, L. Ji, and H. Huang, "Robust Hyperspectral Object Tracking by Exploiting Background-Aware Spectral Information With Band Selection Network," *IEEE Geosci. Remote Sens. Lett.,* Article vol. 19, 2022, Art no. 6013405.

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

[33] E. Ouyang, J. Wu, B. Li, L. Zhao, and W. Hu, "Band Regrouping and Response-Level Fusion for End-to-End Hyperspectral Object Tracking," *IEEE Geosci. Remote Sens. Lett.,* Article vol. 19, 2022, Art no. 6005805.

[34] N. Hien Van, A. Banerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 44-51.

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

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

[37] L. Gao *et al.*, "CBFF-Net: A New Framework for Efficient and Accurate Hyperspectral Object Tracking," *IEEE Trans. Geosci. Remote Sens.,* vol. 61, pp. 1-14, 2023.

[38] Y. Chen *et al.*, "SSTtrack: A Unified Hyperspectral Video Tracking Framework Via Modeling Spectral-Spatial-Temporal Conditions," Available at SSRN: https://ssrn.com/abstract=4860918 or http://dx.doi.org/10.2139/ssrn.4860918, 2024.

[39] Y. Wu, L. Jiao, X. Liu, F. Liu, S. Yang, and L. Li, "Domain Adaptation-aware Transformer for Hyperspectral Object Tracking," *IEEE Trans. Circuits Syst. Video Technol.,* pp. 1-12, 2024.

[40] P. F. Liu, W. Z. Yuan, J. L. Fu, Z. B. Jiang, H. Hayashi, and G. Neubig, "Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *Acm Computing Surveys,* vol. 55, no. 9, Sep 2023, Art no. 195.

[41] S. Chen *et al.*, *AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition*. 2022.

[42] M. L. Jia *et al.*, "Visual Prompt Tuning," in *17th European Conference on Computer Vision (ECCV)*, Tel Aviv, ISRAEL, 2022, vol. 13693, 2022, pp. 709-727.

[43] W. H. Liu, X. Shen, C. M. Pun, X. D. Cun, and Ieee, "Explicit Visual Prompting for Low-Level Structure Segmentations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, CANADA, 2023, 2023, pp. 19434-19445.

[44] K. Sohn *et al.*, "Visual Prompt Tuning for Generative Transfer Learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, CANADA, 2023, 2023, pp. 19840-19851.

[45] Q. Cao, Z. Xu, Y. Chen, C. Ma, and X. Yang, "Domain Prompt Learning with Quaternion Networks." doi: 10.48550/arXiv.2312.08878, 2023.

[46] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song, "Prompting for Multi-Modal Tracking." doi: 10.48550/arXiv.2207.14571, 2022.

[47] T. Xu, X. J. Wu, X. Zhu, and J. Kittler, "Memory Prompt for Spatio-Temporal Transformer Visual Object Tracking," *IEEE Transactions on Artificial Intelligence,* pp. 1-6, 2024.

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

[49] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 40, no. 4, pp. 834-848, Apr 2018.

[50] W. Zhou, F. Sun, Q. Jiang, R. Cong, and J.-N. Hwang, "WaveNet: Wavelet Network With Knowledge Distillation for RGB-T Salient Object Detection," *IEEE Trans. Image Process.,* Article vol. 32, pp. 3027-3039, 2023 2023.

[51] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 7464-7473.

[52] T. Y. Lin *et al.*, "Microsoft COCO: Common Objects in Context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, SWITZERLAND, 2014, vol. 8693, 2014, pp. 740-755.

[53] Y. Wu, J. Lim, and M. H. Yang, "Object Tracking Benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 37, no. 9, pp. 1834-48, Sep 2015.

[54] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the Circulant Structure of Tracking-by-Detection with Kernels," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 7575, 2012, pp. 702-715.

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

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

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

[58] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Learning Spatially Regularized Correlation Filters for Visual Tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015.

[59] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary Learners for Real-Time Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016.

[60] M. Danelljan, G. Hager, F. S. Khan, and M. Felsberg, "Discriminative Scale Space Tracking," *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. 39, no. 8, pp. 1561-1575, Aug 2017.

[61] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning Background-Aware Correlation Filters for Visual Tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 1144-1152.

[62] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan, "Discriminative Correlation Filter Tracker with Channel and Spatial Reliability," *Int. J. Comput. Vis.,* vol. 126, no. 7, pp. 671-688, Jul 2018.

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

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

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

[66] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ECO: Efficient Convolution Operators for Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.

[67] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High Performance Visual Tracking with Siamese Region Proposal Network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.

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

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

[70] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "SiamRPN++: Evolution of Siamese Visual Tracking With Very Deep Networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

[71] L. C. Zhang, A. Gonzalez-Garcia, J. Van De Weijer, M. Danelljan, F. S. Khan, and Ieee, "Learning the Model Update for Siamese Trackers," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Seoul, SOUTH KOREA, 2019, pp. 4009-4018.

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

[73] Q. Wang, L. Zhang, L. Bertinetto, W. Hu, and P. H. S. Torr, "Fast Online Object Tracking and Segmentation: A Unifying Approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.

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

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

[76] Y. D. Xu, Z. Y. Wang, Z. X. Li, Y. Yuan, G. Yu, and I. Assoc Advancement Artificial, "SiamFC plus plus: Towards Robust and Accurate Visual Tracking with Target Estimation Guidelines," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2020, vol. 34, pp. 12549-12556.

[77] D. Y. Guo *et al.*, "Graph Attention Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Electr Network, 2021, pp. 9538-9547.

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

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

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

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

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

[83] X. Chen, H. Peng, D. Wang, H. Lu, and H. Hu, "SeqTrack: Sequence to Sequence Learning for Visual Object Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 14572-14581.

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

[85] Y. Xiao *et al.*, "Local-Global Temporal Difference Learning for Satellite Video Super-Resolution," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 34, no. 4, pp. 2789-2802, 2024.

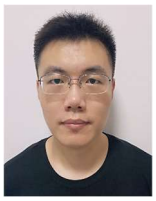
[86] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite Video Super-Resolution via Multiscale Deformable Convolution Alignment and Temporal Grouping Projection," *IEEE Trans. Geosci. Remote Sens.,* Article vol. 60, 2022.

[87] B. Uzkent, A. Rangnekar, and M. J. Hoffman, "Tracking in Aerial Hyperspectral Videos Using Deep Kernelized Correlation Filters," *IEEE Trans. Geosci. Remote Sens.,* Article vol. 57, no. 1, pp. 449-461, Jan 2019.

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

Yuzeng Chen received the B.S. degree in geographic information science from Southwest University of Science and Technology, Mianyang, China, in 2020 and the M.S. degree from Central South University, in 2023, Changsha, China. He is pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan. His research interests include remote sensing/hyperspectral video processing and computer vision. More details can be found at https://yzcu.github.io.

Yuqi Tang (Member, IEEE) received the Ph.D. degree in photogrammetric and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2013. Since 2017, she has been an Associate Professor at the School of Geosciences and Info-Physics, Central South University, Changsha, China. Her research interests include object detection/tracking, land-cover/-use classification, multi-modal remote sensing image change detection, and natural resource monitoring.



Xin Su (Member, IEEE) received the B.S. degree in electronic engineering from Wuhan University, Wuhan, China, in 2008, and the Ph.D. degree in image and signal processing from Télécom ParisTech, Paris, France, in 2015. He was a Post-Doctoral Researcher with the Team SIROCCO, Institut National de Recherche en Informatique et en Automatique, Rennes, France. He is an Assistant Professor with the School of Remote Sensing and Information Engineering, Wuhan University. His research interests include multitemporal remote sensing image processing, multiview image processing, and 3-D video communication.

Jie Li (Member, IEEE) received the B.S. degree in sciences and techniques of remote sensing and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2011 and 2016, respectively. He is an Associate Professor with the School of Geodesy and Geomatics, Wuhan University. His research interests include image quality improvement, image super-resolution reconstruction, data fusion, remote sensing image processing, sparse representation, and deep learning

Yi Xiao (Graduate Student Member, IEEE) received the B.S. degree from the School of Mathematics and Physics, China University of Geosciences, Wuhan, China, in 2020 and the M.S. degree from Wuhan University, Wuhan, China, in 2022. He is pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan. His major research interests include remote sensing image/video processing and computer vision. More details can be found at https://xy-boy.github.io.

Jiang He (Graduate Student Member, IEEE) received the B.S. degree in remote sensing science and technology from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China. His research interests include hyperspectral superresolution, image fusion, quality improvement, remote sensing image processing, and deep learning.

Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

In 2012, he joined the School of Geodesy and Geomatics, Wuhan University, where he is a Professor. He has published more than 90 research articles, including more than 70 peer-reviewed articles in international journals, such as *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE TRANSACTION ON IMAGE PROCESSING, and IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING. His research interests include image reconstruction, remote sensing image processing and application, and data fusion.

Dr. Yuan was a recipient of the Youth Talent Support Program of China in 2019, the Top-Ten Academic Star of Wuhan University in 2011, and the recognition of Best Reviewers of the IEEE GEOSCIENCE AND REMOTE SENSING LETTERS in 2019. In 2014, he received the Hong Kong Scholar Award from the Society of Hong Kong Scholars and the China National Postdoctoral Council. He is an Associate Editor of five international journals and has frequently served as a Referee for more than 40 international journals for remote sensing and image processing.