SENSE: Hyperspectral Video Object Tracker via Fusing Material and Motion Cues

Yuzeng Chen , Qiangqiang Yuan , Yuqi Tang , Yi Xiao , Jiang He , Zhenqi Liu

PII: S1566-2535(24)00173-8
DOI: https://doi.org/10.1016/j.inffus.2024.102395
Reference: INFFUS 102395

To appear in: *Information Fusion*

Received date: 10 December 2023
Revised date: 4 March 2024
Accepted date: 27 March 2024

Please cite this article as: Yuzeng Chen , Qiangqiang Yuan , Yuqi Tang , Yi Xiao , Jiang He , Zhenqi Liu , SENSE: Hyperspectral Video Object Tracker via Fusing Material and Motion Cues, *Information Fusion* (2024), doi: https://doi.org/10.1016/j.inffus.2024.102395

SENSE: Hyperspectral Video Object Tracker via Fusing Material and Motion Cues

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

[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, Hunan 410083, China*
[d] *College of Artificial Intelligence, Southwest University, Chongqing, 400715, China*
** Corresponding author*

ARTICLEINFO

Keywords:
Hyperspectral
Object tracking
Self-expression
False modality fusion
Motion awareness

Highlights:

- A unified hyperspectral video object tracking method with fusing material and motion cues is proposed.
- A spectral-spatial self-expression module is proposed to adaptively obtain complementary false modalities, bridging the band gap issue.
- A cross-false modality fusion module is proposed to aggregate and enhance the differential-common features of false modalities, obtaining robust object representations.
- A motion awareness module is designed that enables continuous tracking of the object in abnormal states.
- Comprehensive experiments and in-depth analysis are conducted to validate the proposed method and provide pre-exploration for future research.

ABSTRACT

Hyperspectral video offers a wealth of material and motion cues about objects. This advantage proves invaluable in addressing the inherent limitations of generic RGB video tracking in complex scenarios such as illumination variation, background clutter, and fast motion. However, existing hyperspectral tracking methods often prioritize the material cue of objects while overlooking the motion cue contained in sequential frames, resulting in unsatisfactory tracking performance, especially in partial or full occlusion. To this end, this article proposes a novel hyperspectral video object tracker via fusing material and motion cues called SENSE that leverages both material and motion cues for hyperspectral object tracking. First, to fully exploit the material cue, we propose a spectral-spatial self-expression (SSSE) module that adaptively converts the hyperspectral image into complementary false modalities while effectively bridging the band gap. Second, we propose a cross-false modality fusion (CFMF) module that aggregates and enhances the differential-common material features derived from false modalities to arouse material awareness for robust object representations. Furthermore, a motion awareness (MA) module is designed, which consists of an awareness selector to determine the reliability of each cue and a motion prediction scheme to handle abnormal states. Extensive experiments are conducted to demonstrate the effectiveness of the proposed method over state-of-the-arts.

1. Introduction

Visual single object tracking, as one of the most fundamental tasks, has found widespread applications in human-machine interaction, traffic analysis, medical image processing, and video surveillance [1-3]. Its primary objective is to

establish the association of an object in a sequence. Significant efforts have been made in object tracking for the red-green-blue (RGB) modality [4-7]. Due to the limited spectral information, RGB-based modality tracking still faces challenges in complex scenarios such as low-light conditions at night, poor visibility caused by fog and haze, similar object appearance, and background clutter [8, 9]. In response to the limitations of a single RGB modality, the integration of multiple modalities such as RGBT (RGB plus Thermal infrared) and RGBD (RGB plus Depth) has emerged as a promising approach to address the aforementioned challenges [10-13]. As depicted in Fig. 1, the depth modality accentuates the intricate three-dimensional structural information of the object, whereas the thermal modality focuses on capturing radiant heat, thereby offering complementary cues to the RGB modality and enhancing tracking performance [14]. However, both RGBD and RGBT modalities necessitate the use of two or more imaging devices. For instance, in RGBT tracking, a combination of CCD and thermal infrared cameras is typically mounted on a platform to concurrently record data. Despite proximity, capturing the same scene with both cameras can be challenging, particularly for small objects at a considerable distance [15, 16]. Consequently, aligning the RGB and Thermal modalities has become standard practice, albeit potentially leading to image distortion issues, as illustrated in Fig. 1(c) and Fig. 1(d).

With the continuous development of imaging devices, hyperspectral (HS) cameras have evolved as powerful tools for capturing comprehensive spectral radiant information of tracked objects [17, 18], equipping trackers with the capability to identify materials [19]. Notably, all spectral bands are captured from the same viewpoint, eliminating the need for cumbersome multi-modality alignment. However, effective utilization of potential modality information in HS video remains an area for further research. Currently, achieving robust HS video tracking encounters several challenges. First, the limited availability of HS video datasets poses a significant obstacle to directly training a robust HS tracker [8, 20]. Second, the

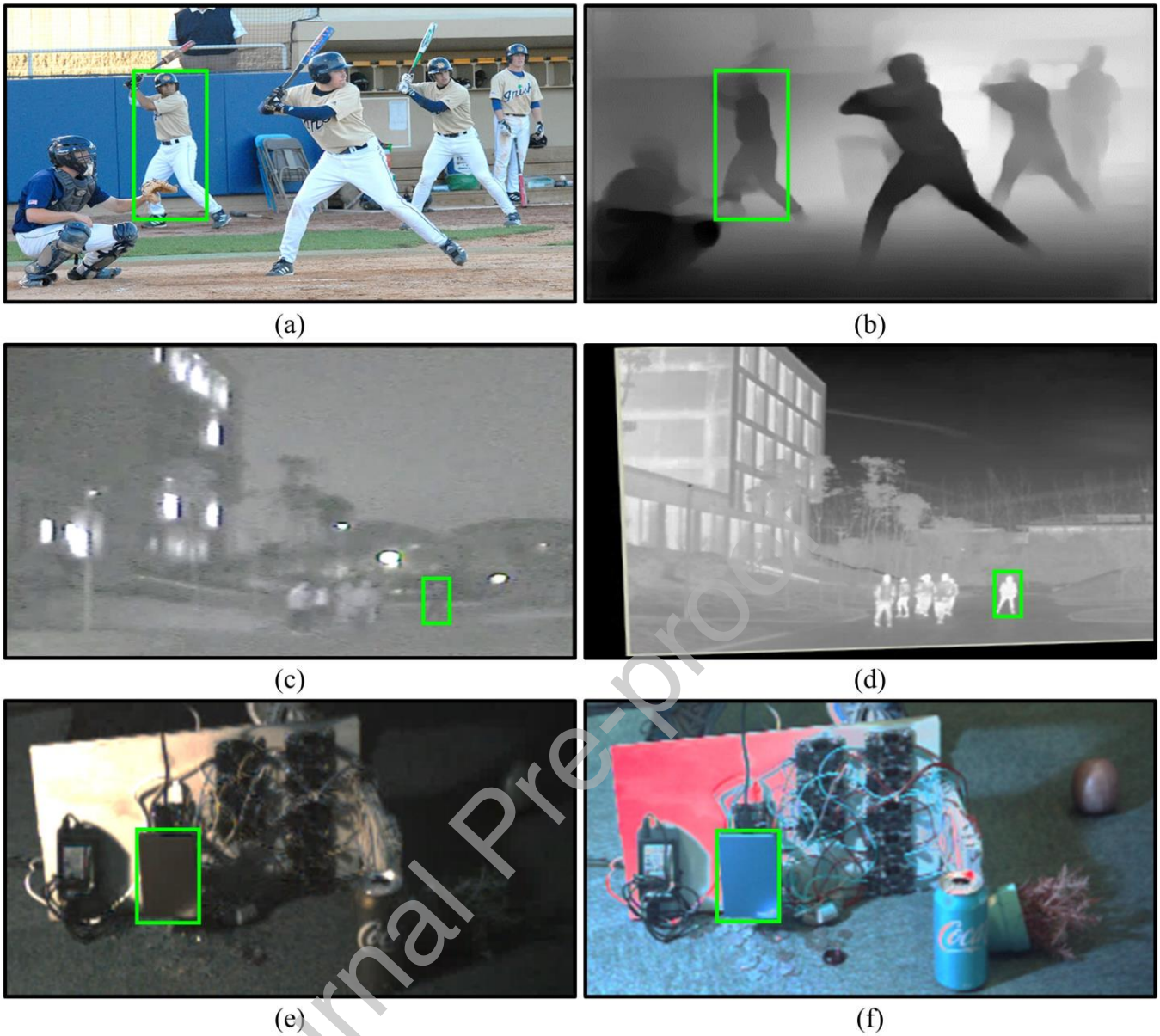


Fig. 1. Visualization of different modalities. (a) RGB modality. (b) Depth modality. (c) RGB modality. (d) Thermal infrared modality. (e) and (f) are false modalities generated from the hyperspectral modality.

disparity of the band numbers between the RGB and HS data (i.e., the band gap) impedes the direct application of pre-trained RGB models to HS tasks [21, 22]. Despite notable endeavors from previous studies [8, 19, 21, 23], they often focus solely on leveraging the material cue of objects and neglect the motion cue present in sequential frames. As a result, their effectiveness may be compromised, particularly in abnormal states such as occlusion (Fig. 2) and fast motion.

Discriminative and invariant features are fundamental for achieving robust HS video tracking, while a suitable representation model determines the ceiling of performance. Previous endeavors such as TSCFW [24], MHT [19], TASSCF [25], and [26] primarily rely on hand-crafted features and inherit correlation filters. Compared to deep features, the limited discriminative capability of hand-crafted features affects the performance of HS trackers [21, 27]. To address the limitation, recent works such as SST-Net [28], BRRF-Net [29], TFTN [30], and SiamOHOT [31] have been developed to exploit deep features for robust object representations, and their performance has been validated. In addition, traditional correlation filters remain difficult to achieve comparable performance against deep Siamese networks, as evidenced by numerous studies in the RGB tracking domain [32-35]. Hence, current HS trackers primarily concentrate on exploring the potential of the deep Siamese network to attain robust performance. As mentioned above, limited HS training samples

present a challenge in

Fig. 2. Visualization of the abnormal state in terms of occlusion. A person is partially occluded (frame #0147) and fully occluded (frame #0154). It is observed that the material cue could be unreliable during occlusion, especially full occlusion. In this case, the motion cue contained in sequential frames can contribute to the continuous tracking. The current frame and object state are marked in the top-left corner, and the yellow box region is enlarged for visibility.

training a generalized deep model. Naturally, large-scale datasets from the RGB tracking domain such as GOT-10K [16], TrackingNet [36], and ImageNet [37] can be reused to pre-train the HS model followed by fine-tuning using HS datasets.

However, the band gap prevents bridging pre-trained RGB models to HS trackers directly [21]. To attenuate this issue, several methods have been proposed such as BAHT [38], BS-SiamRPN [39], SiamHT [40], and [41], which aim to convert the HS image into a three-channel representation through dimension reduction and manual selection. It is worth noting that this fashion inevitably introduces spectral loss and potentially impacts effectiveness [21]. Alternatively, other methods such as BRRF-Net [29], BAE-Net [22], and SST-Net [28] strive to exploit the full band information. They initially transform the HS image into a collection of false color images with equivalent contributions, subsequently employing average ensemble learning to fuse multiple response maps or weak tracking results. This fashion usually achieves better performance. It is important to acknowledge that not all HS bands contribute equally to their formation and downstream tracking tasks [21]. Treating all bands as equal may result in suboptimal outcomes [42]. Recent state-of-the-art (SOTA) approaches such as SEE-Net [21] and SiamBAG [23] have demonstrated the effectiveness of dynamically aggregating weak tracking results based on the contribution of false color images, yielding competitive performance. Additionally, from various perspectives, SOTA HS trackers like TFTN [30] and CBFF-Net [27] have been developed with extensive experiments showcasing their superiority. However, the above efforts often prioritize the material cues at the expense of critical motion cues of the tracked object, especially in complex scenarios where the material cue is unreliable such as the case involving occlusion shown in Fig. 2.

Motivated by the above discussions and analysis, this article aims to fuse both material and motion cues for HS video single object tracking. We propose an end-to-end hyper**s**pectral video object tracker via fusing mat**e**rial and motio**n** cu**es** (SENSE). The key components of SENSE include the spectral-spatial self-expression module, cross-false modality fusion module, and motion awareness module. To bridge the band gap, we propose a spectral-spatial self-expression module, which adaptively partitions the HS image into complementary false color representations with varying contributions. These false color images capture the object's spectral reflectance under diverse wavelengths, similar to multiple false modalities with complementary features, as shown in Fig. 1. Next, these false modalities are fed into a feature extraction module, which utilizes a transferred tracking network pre-trained with RGB data to mitigate the issue posed by limited training samples. Subsequently, we propose a cross-false modality fusion module to aggregate and enhance the differential-common features extracted from false modalities, obtaining robust object representations. Additionally, we design a motion awareness module to determine which cue (i.e., material cue and motion cue) is reliable and predict the final position and scale when the material cue is deemed unreliable.

The primary contributions of this article are summarized as follows.

- A spectral-spatial self-expression (SSSE) module is proposed to capture both spectral and spatial features for effectively

solving the self-expression model. With the SSSE module, the HS image can be dynamically grouped into complementary false modalities with varying contributions, bridging the band gap.

- A cross-false modality fusion (CFMF) module is proposed to aggregate and enhance the differential-common features of false modalities, thereby obtaining robust object representations.
- A motion awareness (MA) module is designed, consisting of an awareness selector to determine which cue (i.e., material and motion) is reliable, as well as a motion prediction scheme to address abnormal states. The SSSE, CFMF, and MA modules are unified into a Siamese network, enabling the SENSE to be both material and motion awareness.

Extensive experiments are conducted to validate the proposed method. The remainder is organized as follows. Sections 2 and 3 provide a review of related work and describe the proposed approach, respectively. Experiments and analysis are presented in Section 4. Finally, Sections 5 and 6 present ablation studies and conclusions, respectively.

2. Review on video tracking methods

In this section, we provide a comprehensive review of object tracking methods in both RGB and HS videos.

2.1. Fusion of RGB and motion for video tracking

Other than the appearance cues, the motion cues are also crucial for action recognition in the tracking paradigm. To improve tracking performance, several studies have been carried out focusing on motion cues such as particle filter, Kalman filter, and optical flow. CPKF [43] contrasts the correlation particle filter method with motion estimation for satellite video object tracking. DOCPF [44] introduces a distractor-occlusion aware correlation particle filter for object tracking in satellite videos. In [45], a novel particle filtering framework is proposed to address template-based visual object tracking probabilistically, and extensive experiments validate the superiority of the proposed framework. BAPF [46] inherits the adaptive particle filter to estimate the proposal and posterior distribution for face detection and tracking in video sequences. In [47], the integration with the Kalman filter and data association techniques is discussed for object representation and localization. DF [7] constructs a dual-mode prediction model to simulate the object's motion pattern and cooperate Kalman filter and non-linear regression to implement object tracking in satellite video. [48] constructs an accurate continuous-discrete extended Kalman filter for flexible and robust radar tracking. In [49], a novel JMMAC tracker using appearance and motion cues is proposed for RGBT video object tracking. For mining the motion information, JMMAC [49] attempts to jointly model the motion patterns of the object and camera by Kalman filter and transformation matrix, respectively. Extensive results demonstrate the superiority of JMMAC in comparison to SOTA algorithms. For the optical flow, it can represent the apparent motion of the brightness patterns and capture information about the magnitude and direction of motion among neighboring frames of a video sequence [50]. Conventional optical flow techniques have been widely used in RGB video object tracking such as VCF [51], MOFT [52], and RAMC [53] Other than the conventional optical flow, recent efforts have explored the potential of deep motion optical flow for object tracking. FlowTrack [54] proposes an end-to-end flow correlation object tracking framework to mine the abundant flow information in sequential video frames. In [55], the authors investigate the impact of deep motion features in a DCF-based tracking framework. They demonstrate that deep optical flow features can provide complementary information to appearance cues for improved tracking performance. In addition, ARTrack [56] presents spatio-temporal prompts to model the sequential evolution of the trajectory propagating motion cues for obtaining more coherent tracking results.

However, previous research has primarily focused on the RGB tracking domain. As above mentioned, RGB-based modality tracking still faces challenges in complex scenarios due to the limited spectral bands [8, 9]. Compared with the above works, the proposed SENSE focuses on the hyperspectral video object tracking domain. It can exploit the abundant material information and model the maneuvering of hyperspectral objects to ensure accurate object position, velocity, and scale, simultaneously. The novelty of SENSE includes the spectral-spatial self-expression module, cross-false modality fusion module, and motion awareness module, which aim to bridge the issues of band gap, limited training samples, and

abnormal states.

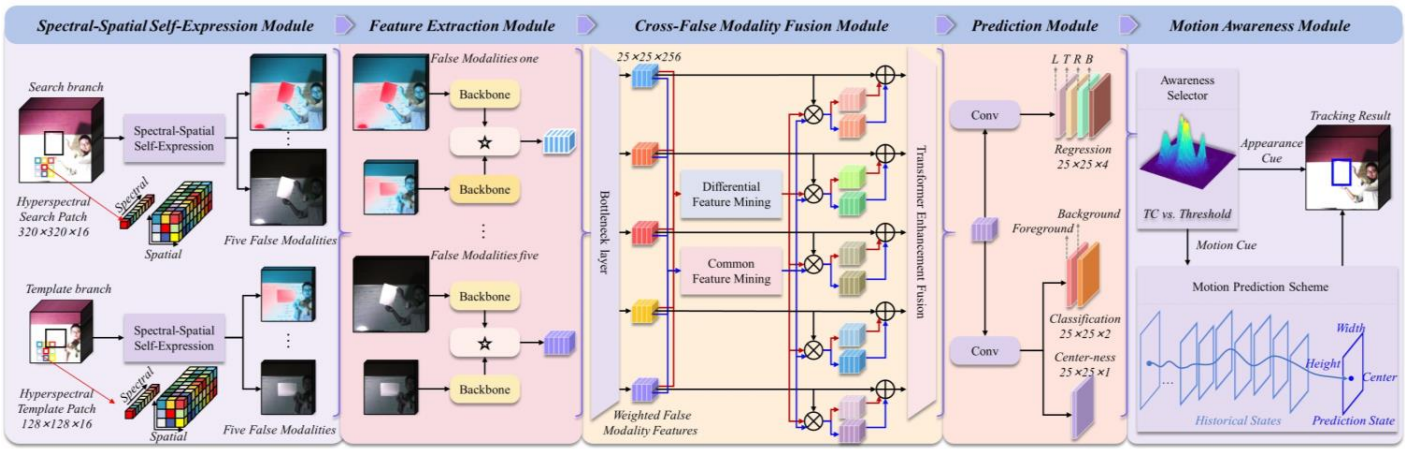


Fig. 3. Overall architecture of the proposed SENSE hyperspectral tracker. The SENSE is a Siamese-based tracker consisting mainly of the proposed spectral-spatial self-expression (SSSE), cross-false modality fusion (CFMF), and motion awareness (MA) modules.

2.2. HS video tracking

The HS data, with its abundance of spectral information, holds inherent advantages over RGB data in object tracking within complex scenarios. Several promising HS trackers have been proposed. Early efforts mainly focus on the generative paradigm, involving the construction of a model to represent the object and the retrieval of a region similar to the model's description. For example, [57] employs the mean shift [58] method for HS object tracking. Inspired by RGB tracking paradigms, recent HS trackers predominantly inherit discriminative correlation filters and Siamese networks. Notably, certain HS trackers, such as MHT [19], TSCFW [24], TASSCF [25], MFI [59], and CNHT [60], are modeled on correlation filters and aim to leverage full band information of HS data. TSCFW [24] explores tensor processing to mitigate spectral differences in homogeneous backgrounds and integrates sparse regularization terms and context-aware information into the correlation filter. Meanwhile, MFI [59] integrates HOG and deep features extracted by the pre-trained VGG-19 [61] network into the correlation filter, yielding robust tracking results. However, these correlation filter-based HS trackers have achieved limited success despite employing hand-crafted and/or deep features.

As mentioned earlier, the robust feature is the basis for reliable tracking, while the appropriate model determines the ceiling of the performance. The Siamese network has gained attention in the field of HS tracking due to its simplicity and discriminative capabilities when compared to traditional correlation filters. Several SOTA HS trackers, such as SEE-Net [21], SiamHYPER [8], SiamOHOT [31], CBFF-Net [27], SiamBAG [23], SiamHT [40], BRRF-Net [29], SSATFN [62], SPIRIT [42], and DT-DBW [63], have incorporated Siamese networks and achieved excellent performance, which also provides a certain prior exploration for our research. For instance, SiamBAG [23] builds a novel band attention grouping-based Siamese framework to address the issue of insufficient training data. To preserve the interaction information between HS bands, CBFF-Net [27] constructs a bidirectional multiple deep feature fusion module and a cross-band group attention module to effectively fuse features. Extensive experiments confirm the effectiveness of these trackers.

However, previous research has primarily focused on exploring the spatial and spectral information, i.e., the material cue, while overlooking the significance of the motion cue contained in HS sequential frames. Yet, the motion cue is particularly important in scenarios where the material cue is unreliable such as occlusion, as shown in Fig.2. Therefore, this article aims to synergize both material and motion cues in a unified framework, proposing an HS video object tracker with material and motion awareness.

3. Method

The proposed HS tracking method will be described, mainly including the overall architecture, spectral-spatial self-expression module, cross-false modality fusion module, motion awareness module, and training and loss.

3.1. Overall architecture

As shown in Fig. 3, the proposed HS tracker consists of five modules: SSSE module, feature extraction module, CFMF module, prediction module, and MA module. The novelty of this paper is that we investigate how to mine abundant physical material cues and motion cues of hyperspectral video objects and fuse these complementary cues into a unified object tracking framework. First, we crop the search patch and template patch from the HS image which are then fed into the spectral-spatial self-expression network of the SSSE module. This module generates complementary false modalities with varying contributions, effectively bridging the band gap. Subsequently, the search and template patches of sorted false modalities are forwarded to the feature extraction module for feature extraction and depth-wise correlation. To obtain a robust object representation, the CFMF is proposed to aggregate the differential-common features that are further enhanced by transformer-like attention. In the prediction module, the material cue of the object is classified and regressed to produce corresponding response maps. These maps are fed into the proposed MA module that comprises an awareness selector and a motion prediction scheme. The awareness selector acts like a tracker switcher that helps determine the reliability of the appearance tracker and motion tracker (i.e., material cue and motion cue). Whereas the motion prediction scheme allows to deal with abnormal states and achieve continuous tracking of the object of interest.

3.2. Spectral-spatial self-expression module

To bridge the band gap, we propose the SSSE module, which adaptively converts an HS patch into complementary false modalities, as shown in Fig. 4. Inspired by the HS self-expression model, the SSSE module inherits the learning-to-optimize fashion to solve the self-expression coefficient matrix, ultimately yielding multiple false modalities. For a given HS video, each frame X is represented as $X = [x_1, x_2, \dots, x_B] \in \mathbb{R}^{D\times B}$, where $D = M \times N$ denotes the number of pixels, B is the band number, and $x_i \in \mathbb{R}^{M\times N}$ represents the i-th band of X. The objective of the SSSE module is to evaluate the band contributions and divide the HS patch into complementary false modalities.

In the field of HS band clustering, the self-expression model is commonly utilized to select a series of HS bands from the original band set, such that each band can be reconstructed by the remaining bands and itself via the self-expression matrix [64]. Correspondingly, the SSSE module aims to solve the coefficient matrix. Considering the Gaussian noise, the self-expression model of all band vectors X can be mathematically expressed by:

$$argmin\|C\|_{1,2}, \mathrm{s.t.}, X = XC + E, \#(1.)$$

where $E \in \mathbb{R}^{D\times B}$ is the residual matrix. $C \in \mathbb{R}^{B\times B}$ denotes the coefficient matrix with $diag(C) = 0$ to eliminate the trivial solution that each band is simply represented by itself. Notably, the i-th row, j-th column, and (i, j)-th element of C are denoted by c^i, c_j, and c_{ij}, respectively. $\|C\|_{1,2} = \sum_{i=1}^{B} \left\|c^i\right\|_2$ indicates the sum of l_2-norm of all row vectors. Moreover, $C \geq 0$ ensures that each nonzero item of c_j denotes the band probability when repressing x_j, and the c_{ij} is constrained by:

$$\sum_{i=1}^{B} c_{ij} = 1, \forall j. \#(2.)$$

To efficiently solve the self-expression model, the Eq. (1) can be written as:

$$argmin_Z \left(\frac{1}{2}\|X - XC\|_F^2 + \frac{\lambda}{2}\|C\|_{1,2}\right), s.t., diag(C) = 0, \#(3.)$$

where $\|\cdot\|_F$ means the Frobenius norm, $\lambda > 0$ is the regularization weight to control the sparsity of C.

To circumvent expensive iterative optimization processes, the learning-to-optimize fashion has been validated by [21]. However, [21] primarily emphasizes spectral dimension features and neglects to consider the spatial relationship between

the object pixels and neighboring pixels. Therefore, we absorb the learning-to-optimize fashion to train a deep HS optimizer (SSSE module) to compute the coefficient matrix C with consideration of both spectral and spatial dimensions. SSSE module consists of an encoder and a decoder connected in cascade, as shown in Fig. 4. To be specific, the encoder comprises a spectral network to first extract spectral features of each pixel and a spatial network to capture spatial features from

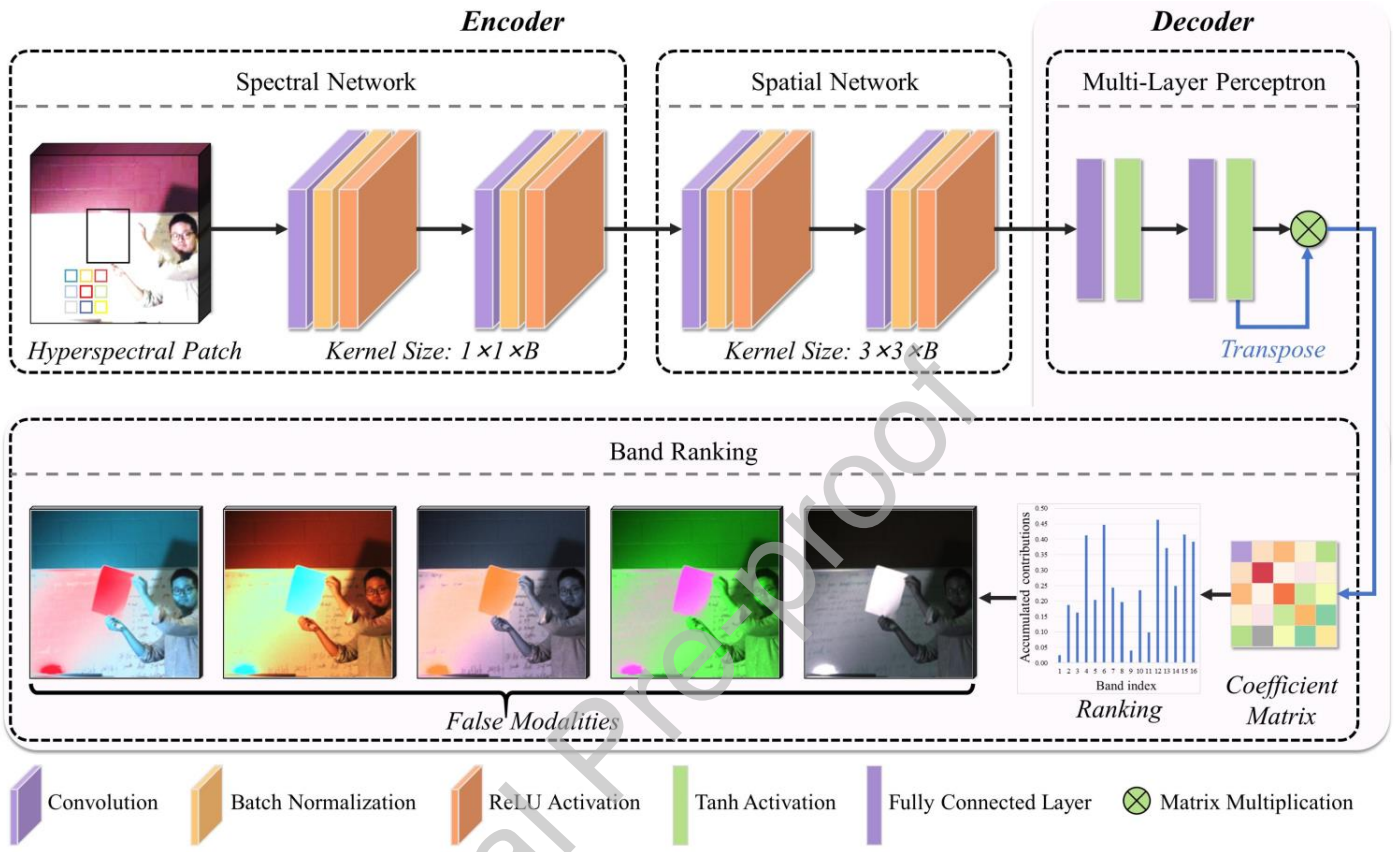


Fig. 4. Structure of the proposed SSSE module. The input is an HS patch while the output is false modalities. The encoder comprises a spectral network to capture spectral features, and a spatial network to provide complementary spatial features. The decoder consists of a multi-layered perceptron for generating a coefficient matrix and a band ranking for adaptively grouping $K = int(B/3)$ false modalities. B denotes the number of channels.

neighboring pixels. On the other hand, the decoder comprises a multi-layer perceptron (MLP) for obtaining the coefficient matrix and a band ranking mechanism for generating complementary false modalities. The encoder is capable of capturing spectral and spatial features from the HS patch, which can be combined with the decoder and downstream tasks for end-to-end training. Detailed descriptions of the encoder and decoder will be presented below.

Concerning the encoder, the spectral network (Fig. 4) emphasizes the channel dimension information and highlights the spectral features of each HS pixel, which can be derived by:

$$X_{spe} = \delta\left(\mathcal{B}\left(Conv_2\left(\delta\left(\mathcal{B}(Conv_1(X))\right)\right)\right)\right), \#(4.)$$

where $X_{spe} \in \mathbb{R}^{M\times N\times 4B}$ represents the extracted spectral feature, $X = [x_1, x_2, \dots, x_B] \in \mathbb{R}^{M\times N\times B}$ denotes the HS patch reshaped from the aforementioned $X \in \mathbb{R}^{D\times B}$, $Conv_1$ and $Conv_2$ are two normal convolutions with kernel size and stride of 1×1 and 1, $\mathcal{B}$ is the batch normalization (BN), and δ is the Rectified Linear Unit (ReLU) activation.

For further digging out the spatial information of neighboring pixels, X_{spe} is fed to the spatial network by:

$$X_{spa} = \delta\left(\mathcal{B}\left(Conv_4\left(\delta\left(\mathcal{B}\left(Conv_3(X_{spe})\right)\right)\right)\right)\right), \#(5.)$$

where $X_{spa} \in \mathbb{R}^{M\times N\times 2B}$ denotes the resulting spectral-spatial features, $Conv_3$ and $Conv_4$ are two normal convolutions with a kernel size of 3×3, a stride of 1, and a padding of 1 to maintain the feature size. The encoder captures spectral-spatial features from the channel and spatial dimensions.

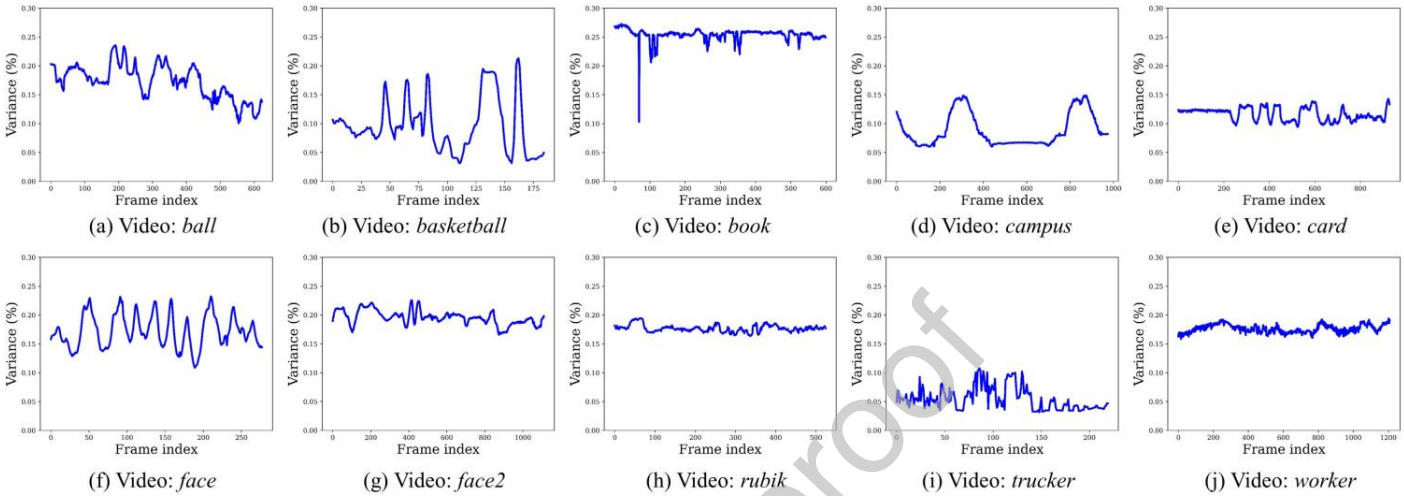


Fig. 5. Standard deviation of the accumulated contributions $Z = [z_1, z_2, \cdots, z_B] \in \mathbb{R}^{B\times 1}$ in sample videos. In each subplot, the horizontal axis denotes the frame index and the vertical axis is the variation, i.e., the standard deviation.

For the decoder, an MLP is first applied to decode the encoded spectral-spatial features for generating the coefficient matrix. Subsequently, the accumulated contributions of each HS band are computed to facilitate the adaptive grouping of HS patches into complementary false modalities. Specifically, the decoding process is implemented by the MLP layer to ensure efficiency and full interaction with deep semantic information:

$$Y = \sigma\left(FC_2\left(\sigma\left(FC_1(X'_{spa})\right)\right)\right), \#(6.)$$

where X'_{spa} represents the flattened result of X_{spa}, FC_1 and FC_2 denote two fully connected layers, σ is the Hyperbolic Tangent (Tanh) activation and $Y \in \mathbb{R}^{(M\times N)\times B}$ is the attention matrix of X'_{spa}. We then obtain the self-expression coefficient matrix $C = Y^T Y$. Matrix $C = [c_1, c_2, \ldots, c_B] \in \mathbb{R}^{B\times B}$ inherently reveals the interdependence among spectral bands. For an arbitrary spectral band x_i, it can be reconstructed by embedding c_i from remaining (including itself) bands. Intuitively, the more important a band is, the larger the coefficient tends to be provided by its embedding c_i. Hence, we can select a subset of important bands by matrix C. Thereafter, all bands are ranked in descending order according to their accumulative contributions. As defined earlier, c^i (the i-th row of C) signifies the contribution of the i-th band to the reconstruction. While c_j (the j-th column of C) stands for the coefficient for reconstructing the j-th band using the remaining HS bands. To acquire the accumulative contribution, we first normalize C along the column direction by:

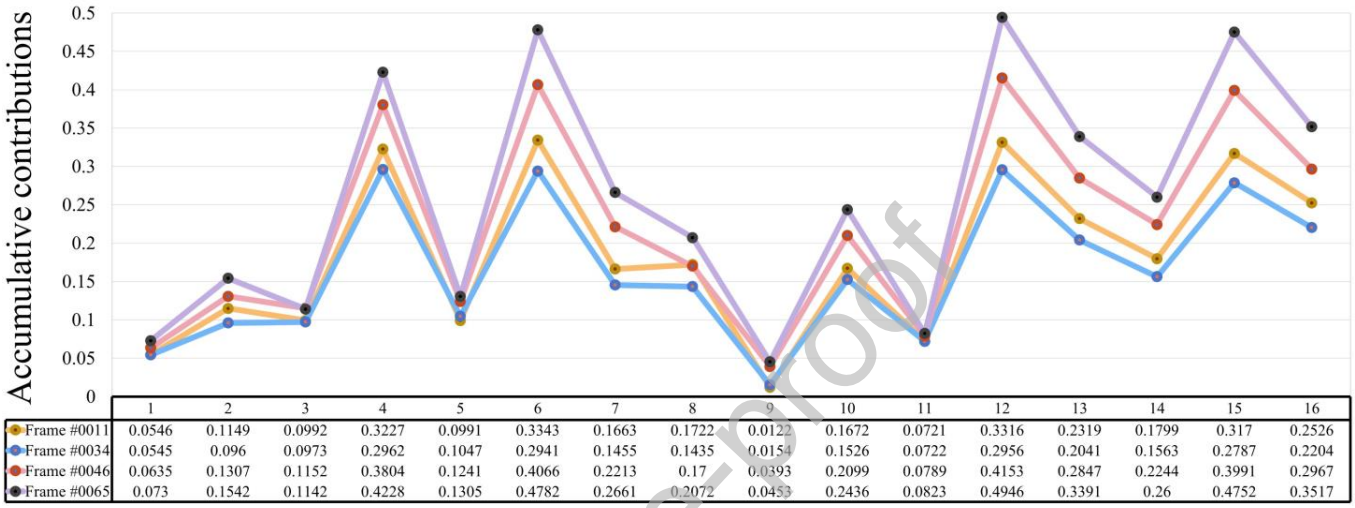
$$\check{c}_j = \frac{|c_{ij}|}{\|c_j\|_2}, \forall i, \#(7.)$$

where $\check{c}_j$ denotes the normalization result of the j-th column of matrix C. Then, the accumulative contribution is obtained by:

$$v_i = \|\check{c}^i\|_1, \#(8.)$$

where $\check{c}^i$ denotes the normalization result of the i-th row of matrix C, $v_i \in \mathbb{R}^{1\times 1}$ is the accumulative contribution of the i-th band, $Z = [z_1, z_2, \cdots, z_B] \in \mathbb{R}^{B\times 1}$ indicates the desired cumulative contributions for each HS band. Fig. 5 shows the standard deviation of $Z = [z_1, z_2, \cdots, z_B]$. It is noticed that the standard deviation is always greater than zero in each video, indicating the acquisition of a set of false modalities with varying Z. In addition, the standard deviation varies across different videos and frames, implying the adaptive variation of Z. Fig. 6 displays the accumulated contributions Z of each band in four sample frames, while Fig. 7 depicts the generated complementary false modalities of sample frames. As can be observed, these false modalities exhibit fewer continuous bands with relatively sparse distribution.

Upon the fact that neighboring bands are highly correlated, less redundancy would be contained in generated false



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Frame #0011	0.0546	0.1149	0.0992	0.3227	0.0991	0.3343	0.1663	0.1722	0.0122	0.1672	0.0721	0.3316	0.2319	0.1799	0.317	0.2526
Frame #0034	0.0545	0.096	0.0973	0.2962	0.1047	0.2941	0.1455	0.1435	0.0154	0.1526	0.0722	0.2956	0.2041	0.1563	0.2787	0.2204
Frame #0046	0.0635	0.1307	0.1152	0.3804	0.1241	0.4066	0.2213	0.17	0.0393	0.2099	0.0789	0.4153	0.2847	0.2244	0.3991	0.2967
Frame #0065	0.073	0.1542	0.1142	0.4228	0.1305	0.4782	0.2661	0.2072	0.0453	0.2436	0.0823	0.4946	0.3391	0.26	0.4752	0.3517

Fig. 6. Visualization of the accumulated contributions $Z = [z_1, z_2, \cdots, z_B] \in \mathbb{R}^{B\times 1}$ in four sample frames (i.e., #0011, #0034, #0046, and #0065). The horizontal axis indicates the band index with a total of 16 bands.

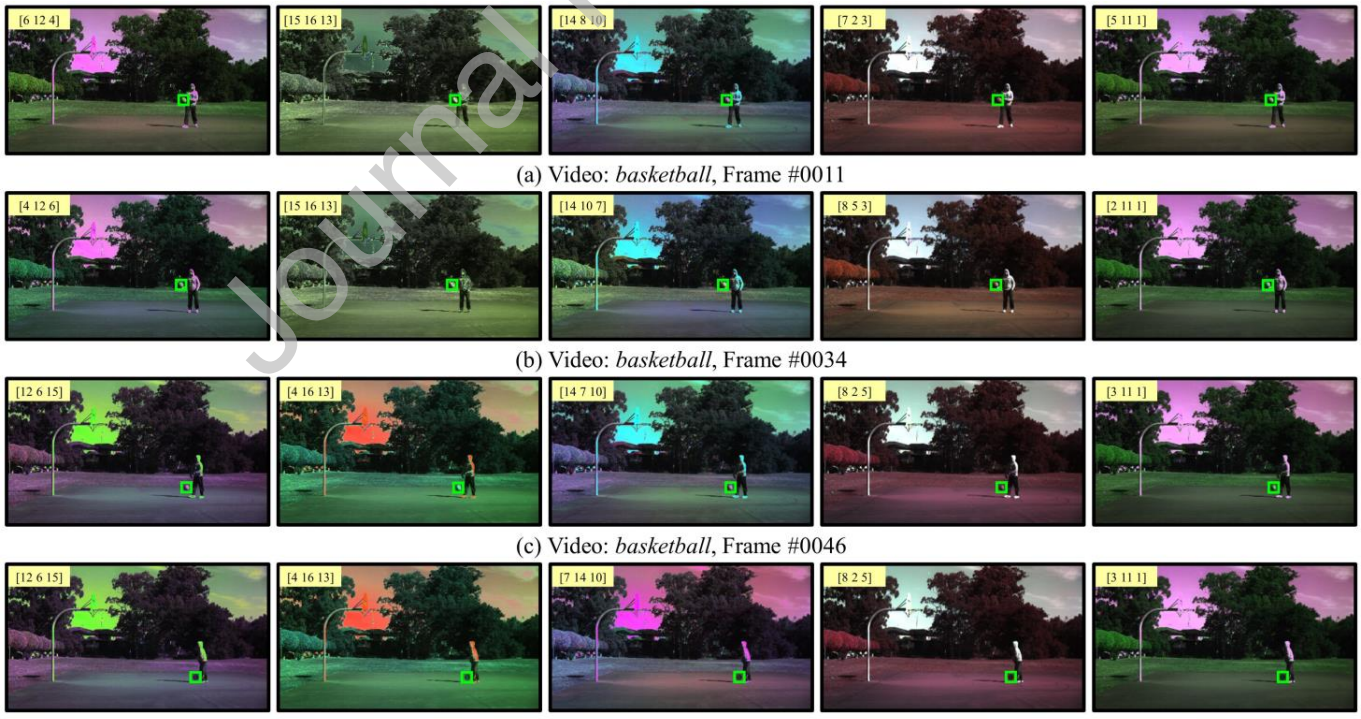


Fig. 7. Visualization of generated false modalities of four sample frames (i.e., #0011, #0034, #0046, and #0065). The band index and tracked object are marked in the top-left corner and the green bounding box, respectively. As can be found, the SSSE module is capable of adaptively dividing the HS image into complementary false modalities with less redundancy.

modalities. Based on $Z \in \mathbb{R}^{B\times 1}$, we first rank all bands and then create $K = int(B/3)$ false modalities $Q = [q_1, q_2, \cdots, q_K]$ by grouping adjacent bands, where $q_i \in \mathbb{R}^{M\times N\times 3}$ represents the i-th false modality. The contribution of each false modality is derived by summing the accumulative contributions of all bands and dividing by the number of bands. Finally, we can get all false modalities $Q = [q_1, q_2, \cdots, q_K]$ (Fig. 7) and their contributions $W = [w_1, w_2, \cdots, w_K] \in \mathbb{R}^{K\times 1}$.

Notably, the proposed SSSE module is trained in an end-to-end fashion and propagates information from the downstream tracking task to the learning of band contributions. As a result, the information learned from the tracking task is also propagated backward to facilitate the evaluation of band contributions, which allows the SSSE module to adapt to the tracking task rather than just benefiting from the self-expression model.

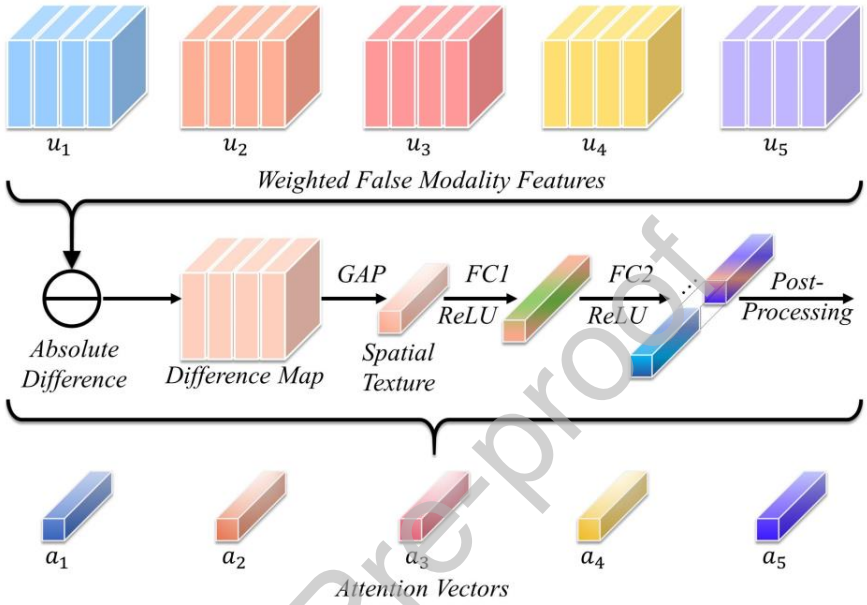


Fig. 8. Structure of the proposed differential feature mining network.

3.3. Cross-false modality fusion module

HS images record continuous spectral information rather than the monochromatic or color information of objects. Spectral information provides a detailed characterization of the material composition, enhancing the capability to discriminate among objects. In the SSSE module, the HS image is adaptively partitioned into false modalities that describe the complementary (i.e., differential and common) material features from distinct spectral perspectives. Therefore, a CFMF module (Fig. 3) is proposed to aggregate and enhance these false modality features. The details of the CFMF module will be presented including the differential-common feature aggregation and transformer enhancement fusion.

3.3.1. Differential-common feature aggregation

In this section, the differential and common features across false modalities are adaptively extracted and aggregated. As mentioned above, the SSSE module can group HS image $X = [x_1, x_2, \ldots, x_B] \in \mathbb{R}^{M\times N\times B}$ into false modalities $Q = [q_1, q_2, \cdots, q_K]$ with the contribution of $W = [w_1, w_2, \cdots, w_K] \in \mathbb{R}^{K\times 1}$. W not only records the significance of false modalities in composing the HS data but also reflects their influence on the downstream task. Therefore, W can be regarded as the initial weight for fusing these false modality features, enabling effective information extraction from limited HS data. In the beginning, the original false modalities

are fed to the feature extraction module to generate semantic features, which are input into the CFMF module. After a bottleneck layer (Fig. 3), we can obtain the initial weighted features of false modalities:

$$U = \frac{w_i \varphi(q_i)}{\sum_{i=1}^{K} w_i}, i = 1, 2, \ldots, K, \#(9.)$$

where $q_i \in \mathbb{R}^{M\times N\times 3}$ is the i-th false modality, w_i denotes the contribution of q_i, $\varphi(\cdot)$ is the feature extraction operation, K is the number of false modalities and $U = [u_1, u_2, \cdots, u_K]$ is the initial weighted result of the CFMF module. $u_i \in \mathbb{R}^{M'\times N'\times K'}$ is the i-th feature of U corresponding to q_i. Specifically, a 16-band HS image is divided into five false modalities and weighted features, i.e., $K = 5$ and $U = [u_1, u_2, u_3, u_4, u_5]$.

As discussed above, these false modalities describe the differential and common material features, which are beneficial for robust object representations. To dig out differential modality features, the difference map (Fig. 8) is first obtained by performing the absolute difference operation:

$$L = \left|\left||u_1 - u_2| - u_3\right| - u_4\right| - u_5\Big|, \#(10.)$$

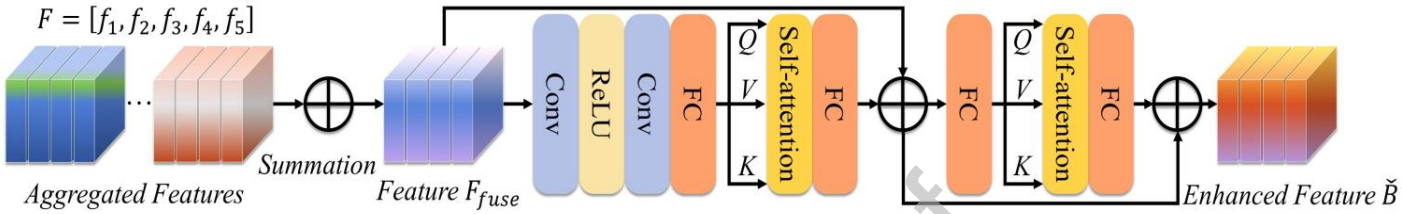


Fig. 9. Structure of the proposed transformer enhancement fusion network.

where $|\cdot - \cdot|$ stands for the element-wise absolute difference operation to guarantee that difference features are non-negative and meaningful, and $L \in \mathbb{R}^{M'\times N'\times K'}$ denotes the difference map. To reveal inter-channel dependencies and embed the global information of the difference feature, the channel attention is introduced, as shown in Fig. 8. More concretely, the spatial context descriptor G is first generated by squeezing the spatial dimensions of difference features using global average pooling (GAP):

$$G = \frac{1}{M' \times N'}\sum_{i=1}^{M'}\sum_{j=1}^{N'} L(i,j), \#(11.)$$

where M' and N' indicate the size of L. $G \in \mathbb{R}^{K'\times 1}$ represents the spatial texture descriptor that is passed forward to two fully connected layers with ReLU activation to create the attention vector $A = [a_1, a_2, a_3, a_4, a_5]$. $a_i \in \mathbb{R}^{K'\times 1}$ stands for the channel weight of the i-th false modality feature u_i, as presented in Fig. 8. Mathematically, the operation process can be represented by:

$$A = \xi\left(\delta\left(FC_4\left(\delta\left(FC_3(G)\right)\right)\right)\right), \#(12.)$$

where FC_3 and FC_4 are two fully connected layers, and ξ integrates the post-processing operations including the reshape, softmax, and chunk operations. The i-th differential modality feature $u_i' \in \mathbb{R}^{M'\times N'\times K'}$ is computed by $u_i' \otimes u_i$, where $\otimes$ denotes the element-wise multiplication operator. Finally, we can obtain five sets of differential modality features $U^d = [u_1^d, u_2^d, u_3^d, u_4^d, u_5^d]$.

The differential modality feature U^d reflects the difference information while ignoring the common information among modalities. To this end, the common modality features are further exploited as complementary information for robust object representations.

Let $L' \in \mathbb{R}^{M'\times N'\times K'}$ denote the commonality map that is obtained by performing the addition operation between U:

$$L' = u_1 \oplus u_2 \oplus u_3 \oplus u_4 \oplus u_5, \#(13.)$$

where $\oplus$ denotes the element-wise summation.

After obtaining L', the subsequent operations are similar to the acquisition of differential modality features, and we can obtain the common modality features $U^c = [u_1^c, u_2^c, u_3^c, u_4^c, u_5^c]$.

Inspired by the concept of residuals, the complementary (i.e., differential and common) features are added to the original

features, thereby enhancing the stability of the network and being capable of obtaining remote context representations. The process is achieved by:

$$F = U^d \oplus U^c \oplus U, \#(14.)$$

where $F = [f_1, f_2, f_3, f_4, f_5]$ denotes the aggregated differential-common feature with $f_i = u_i^d \oplus u_i^c \oplus u_i$.

3.3.2. Transformer enhancement fusion

Aggregating and enhancing different modal information presents a critical challenge when working with complementary false modalities. To address this challenge, the differential-common feature aggregation part has implemented adaptive aggregation of false modalities. Furthermore, we strive to achieve self-enhancement of f_{fuse} to obtain more robust object representations. To accomplish this goal, the transformer-like attention mechanism is introduced in the CFMF module. Fig. 9 shows the structure of the proposed transformer enhancement fusion. It takes the aggregated feature $F = [f_1, f_2, f_3, f_4, f_5]$ as input and performs summation by $F_{fuse} = f_1 \oplus f_2 \oplus f_3 \oplus f_4 \oplus f_5 \in \mathbb{R}^{M' \times N' \times K'}$. Then, F_{fuse} are converted into three vectors including query, key, and value. The transformer self-attention weight matrix is produced from the query and key and is then multiplied by the value. Following a fully connected layer, the output is summed with the original input. The transformer self-attention is repeated to get enhanced features. Formulaically, the operation is expressed as

$$\check{B} = FC_6\left(Attention_1\left(FC_5\left(Conv_6\left(\delta\left(Conv_5\left(F_{fuse}\right)\right)\right)\right)\right)\right) \oplus F_{fuse}, \#(15.)$$

$$B = FC_8\left(Attention_2\left(FC_7\left(\check{B}\right)\right)\right) \oplus \check{B}, \#(16.)$$

where B and $\check{B}$ denote the intermediate result and final enhanced features, respectively. FC_5 and FC_7 are responsible for dimension reduction and generation of the query, key, and value. FC_6 and FC_8 is used to raise the feature dimension. $Attention_1$ and $Attention_2$ refer to two transformer self-attentions. For each transformer-like attention, it takes the query, key, and value as input and the weighted sum of values as output. The weight assigned to the value is derived by performing the softmax operation of the scaled dot products between the query and key. The attention is formulated as

$$Attention(\mathcal{Q}, \mathcal{K}, \mathcal{V}) = softmax\left(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{d_k}}\right)\mathcal{V}, \#(17.)$$

where matrices $\mathcal{Q}$, $\mathcal{K}$, and $\mathcal{V}$ denote query, key of dimension d_k, and value, respectively. In summary, the CFMF module can achieve differential-common feature aggregation and transformer enhancement fusion.

3.4. Motion awareness module

The motion cue plays a vital role in HS video object tracking, especially when the material cue is unreliable, as shown in Fig. 3. At this point, we can model the maneuvering of the tracked object depending on the historical state [65]. Towards this end, the simultaneous use of appearance and motion cues is explored for HS video object tracking. For each frame of the video, we use motion cues to predict the initial position of the object followed by integrating appearance cues to search for the accurate position. However, the results are unsatisfactory compared to the fashion of using the awareness selector shown in Fig. 3. This may be because appearance cues are more dominant than motion cues and usually play a prominent role in anomalous states such as occlusion. However, object anomalies are usually less. Therefore, we propose an awareness selector to determine the reliability of material and motion cues as done in [49] and [66]. To continuously track an object under abnormal states such as an occlusion. An HS tracker is expected to possess the following capabilities.

- Awareness of interference: A tracker needs to be aware of the occurrence of object interference.

- Handling of interference: When the object is in an abnormal state, ensure that the tracker does not lose it.
- Awareness of the end of interference: A tracker can be aware of the end of object interference.

However, current HS trackers focus little on the motion cue contained in sequential frames. In this work, we propose to combine the material and motion cues into a unified tracking framework and design a motion awareness module to handle abnormal states. The motion awareness module comprises an awareness selector to determine which cue (i.e., material or motion) is reliable and a motion prediction scheme to predict the position, velocity, and scale of the tracked object, handling abnormal states.

3.4.1. Awareness selector

In the proposed method, we jointly exploit the material and motion cues for tracking. Empirically, the material cue is more discriminative than the motion cue in most cases. Whereas, it is difficult to use the material cue for localization when

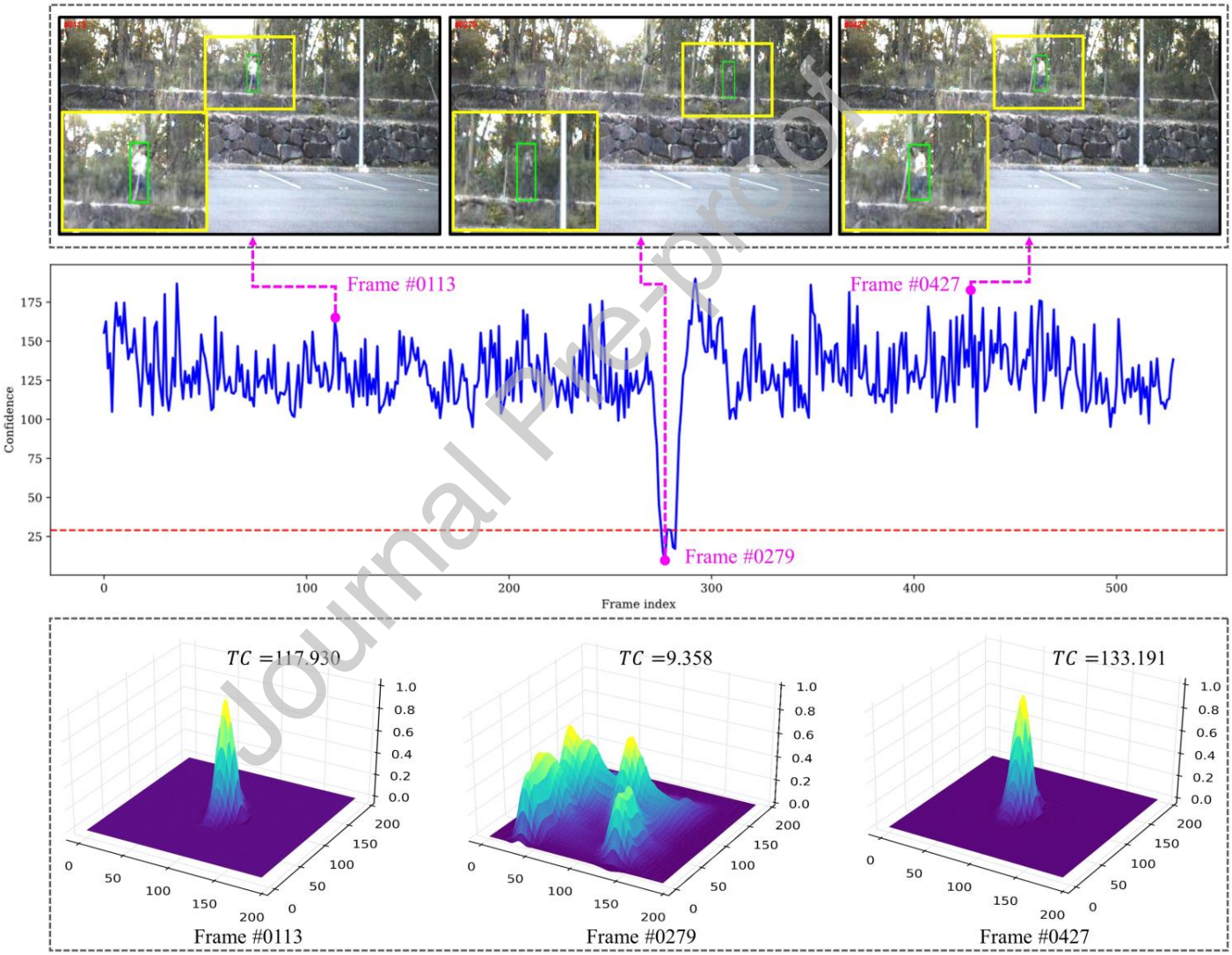


Fig. 10. Visualization of TC illustrated by the *forest* video. The larger the TC, the more reliable the material cue. The first row shows the video frame with the yellow-boxed region zoomed in. The second row displays the confidence curve with the red dashed line indicating the threshold τ. The third line exhibits corresponding response maps. As can be observed, the material cue is unreliable at frame #0279 due to the occlusion. In this case, the response map tends to be flat and multi-peaked distribution, resulting in a low TC value. Therefore, we can determine whether the object is in an abnormal state according to TC.

the object is in an abnormal state (Fig. 2). Therefore, we design the awareness selector to determine which cue is more

reliable and suitable for tracking the current object. The awareness selector is achieved by evaluating the proposed tracking confidence TC of the response map obtained from the material cue

$$TC = \frac{RM_{max}|RM_{max} - RM_{min}|^2}{mean\left(\sum_{i=1}^{H}(RM_i - RM_{min})^2\right)}, \#(18.)$$

where RM_{max} and RM_{min} indicate the maximum and minimum values of the response map RM, respectively, and H is the number of pixels of RM. This is because peaks and fluctuations of response maps can reveal the confidence of tracking results. When the detected result is well-matched to the correct object, the ideal response map should have only one sharp peak, and all other areas should be smooth. The sharper the correlation peak, the higher the localization precision. Otherwise, the entire response map will fluctuate violently with significant pattern differences from the normal response map. Inspired by this, the peak value and average peak-to-correlation energy of the response map are proposed to explore a tracking confidence feedback mechanism. Fig. 10 shows the TC value during tracking, in which the larger the TC, the more reliable the material cue. It can be observed that TC decreases significantly when the material cue is unreliable. Therefore, we can determine whether to activate the motion prediction scheme or not by comparing TC with the threshold τ.

3.4.1. Motion prediction scheme

The motion prediction scheme will be activated when TC is smaller than τ. Since the sampling interval between every two frames is short, we can assume that the object exhibits uniform linear motion and maintains a constant scale variation. Guided by the Kalman filter [67], a motion prediction scheme is proposed to simultaneously predict the center position, velocity, and scale of the tracked object. Let $S_k = \left[x_k, y_k, w_k, h_k, v_k^x, v_k^y, v_k^w, v_k^h\right]^T$ stand for the state vector at frame k, where (x_k, y_k) and (w_k, h_k) are the center coordinates and the width-height, and $\left(v_k^x, v_k^y\right)$ and $\left(v_k^w, v_k^h\right)$ denote the object's velocity of motion and scale variation in horizontal and vertical directions. The proposed motion prediction scheme involves two stages, i.e., prediction and updating. In the prediction stage, the state and error transformation equations are

$$\hat{S}_{\bar{k}} = M\hat{S}_{k-1} + Du_{k-1}, \#(19.)$$

$$E_{\bar{k}} = ME_{k-1}M^T + Q_k, \#(20.)$$

where $\hat{S}_{\bar{k}}$ is a prior state estimation at frame k, $\hat{S}_{k-1}$ is the posterior state estimation at frame $k-1$, D indicates the control matrix, u_{k-1} is the control vector with covariance matrix Q, $E_{\bar{k}}$ is the prior estimation of the error matrix, and M denotes the state transformation matrix with

$$M = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \#(21.)$$

The observation equation can be expressed by

$$Z_k = HS_k + V_k, \#(22.)$$

where Z_k denotes the observation vector, and V_k is noise with covariance matrix R. $H \in \mathbb{R}^{4\times 8}$ is the observation matrix

$$H = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}. \#(23.)$$

In the updating stage, we apply the observation vector Z_k to correct the prior estimation $\hat{S}_{\bar{k}}$ with errors, obtaining the posterior estimate $\hat{S}_k$ at frame k. The main process can be expressed as follows

$$K_k = \frac{E_{\bar{k}} H^T}{H E_{\bar{k}} H^T + R_k}, \#(24.)$$

$$\hat{S}_k = \hat{S}_{\bar{k}} + K_k\left(Z_k - H\hat{S}_{\bar{k}}\right), \#(25.)$$

$$E_k = (I - K_k H) E_{\bar{k}}, \#(26.)$$

where $\hat{S}_k$ is the desired posteriori state vector corrected by Z_k, I is the identity matrix, and K_k stands for the gain matrix. It is worth noting that the motion prediction scheme is not updated when the motion cue is selected to determine final outputs due to the absence of actual measurements. Otherwise, the material cue will be considered as actual measurements to update the scheme, ensuring accurate object position, velocity, and scale.

3.5. Training and loss

3.5.1. Training

Due to the limited HS training samples, the effectiveness of the designed network would be unsatisfactory when initialized randomly. For this reason, the feature extraction and prediction modules are initialized with parameters provided by [32] and remain frozen. While the spatial-spectral self-expression module and cross-false modality fusion module are trained from scratch, without a pre-trained model, and the motion awareness module is executed in the inference process. The sizes of the input search and template patches are set to 255×255×16 and 127×127×16, respectively.

3.5.2. Loss

The proposed method is trained with multi-task loss, as follows

$$\mathcal{L}_{total} = \gamma_1 \mathcal{L}_{rec} + \gamma_2 \mathcal{L}_{reg} + \gamma_3 \mathcal{L}_{cls} + \gamma_4 \mathcal{L}_{cen}, \#(27.)$$

where $\mathcal{L}_{rec}$ stands for the reconstruction loss, $\mathcal{L}_{reg}$ denotes the regression loss of the bounding box, $\mathcal{L}_{cls}$ is the cross-entropy loss for classification, $\mathcal{L}_{cen}$ is the center-ness loss for estimating the localization quality.

Constants γ_1, γ_2, γ_3, and γ_4 weight the reconstruction loss, regression loss, classification loss, and center-ness loss. For the $\mathcal{L}_{rec}$, it is the average loss of the template loss and the search loss

$$\mathcal{L}_{rec} = \frac{1}{2}(\mathcal{L}_{rec}^{tem} + \mathcal{L}_{rec}^{sea}), \#(28.)$$

where $\mathcal{L}_{rec}^{tem}$ and $\mathcal{L}_{rec}^{sea}$ stand for the loss of the self-expression model for the template and the search region.

For instance, the loss $\mathcal{L}_{rec}^{sea}$ is measured by the mean squared error

$$\mathcal{L}_{rec}^{sea} = \sum_{i=1}^{B} \| \hat{x}_i - x_i \|^2, \#(29.)$$

where B is the number of bands, and $\hat{x}_i$ and x_i are the i-th band of the predicted reconstruction result and ground truth.

As shown in Fig. 3, the tracking task is divided into two subtasks, i.e., the classification task that predicts the category of each location and the regression task that computes the object bounding box for that location. Specifically, the classification branch and regression branch output the classification feature map $A_{cls} \in \mathbb{R}^{w \times h \times 2}$ and regression feature maps $A_{reg} \in \mathbb{R}^{w \times h \times 4}$, where w and h refer to the width and height of the extracted feature maps, respectively. Furthermore, each point $(i, j, :)$ in A_{cls} is a 2D vector representing the foreground and background scores at the corresponding position of the search region.

Similarly, each point $(i, j, :)$ in A_{cls} is a 4D vector $t(i, j) = (l, t, r, b)$ indicating the distance from the corresponding position to the four sides (i.e., left, top, right, and bottom) of the box in the search region. Thus, we can train the tracker

using cross-entropy loss for classification and intersection over union (IoU) loss for regression.

Let (x^{lt}, y^{lt}) and (x^{rb}, y^{rb}) represent the upper-left and right-bottom corner positions of the ground truth, respectively. And the (x, y) is the corresponding position of points (i, j) in the search region. Then, the regression object $o_{(i,j)}$ at A_{reg} is obtained by

$$\begin{aligned} \hat{o}^{l}_{(i,j)} &= \hat{l} = x - x^{lt}, \quad \hat{o}^{t}_{(i,j)} = \hat{t} = y - y^{lt}, \\ \hat{o}^{r}_{(i,j)} &= \hat{r} = x^{rb} - x, \quad \hat{o}^{b}_{(i,j)} = \hat{b} = y^{rb} - y, \end{aligned} \#(30.)$$

where symbol $\hat{\cdot}$ denotes the predicted value. After obtaining $\hat{o}_{(i,j)}$, we could calculate the IoU between the predicted box and ground truth and get the regression loss $\mathcal{L}_{reg}$

$$\mathcal{L}_{reg} = \frac{1}{\sum \mathbb{I}(\hat{o}_{(i,j)})} \sum_{i=1}^{w} \sum_{j=1}^{h} \mathbb{I}(\hat{o}_{(i,j)}) \mathcal{L}_{iou}(A_{reg}(i,j,:), \hat{o}_{(i,j)}), \#(31.)$$

where $\mathcal{L}_{iou}$ and $\mathbb{I}(\cdot)$ denote the IoU loss and indicator random variable that is

$$\mathbb{I}(\hat{o}_{(i,j)}) = \begin{cases} 1 & if\ \forall(\hat{l}, \hat{t}, \hat{r}, \hat{b}) > 0 \\ 0 & otherwise. \end{cases} \#(32.)$$

The object's center plays an important role in determining the bounding box, and the farther the prediction location is from the center, the lower the quality of the resulting prediction box [32]. To attenuate this problem, a center-ness algorithm is imposed to improve the localization quality by producing the center-ness feature map $A_{cen} \in \mathbb{R}^{w \times h \times 1}$. Each point $(i, j, :)$ in A_{cen} records the center-ness score $CS(i, j)$ of the corresponding position in the search, and $CS(i, j)$ is computed by

$$CS(i,j) = \mathbb{I}(\hat{o}_{(i,j)}) * \sqrt{\frac{min(\hat{l}, \hat{r})}{max(\hat{l}, \hat{r})} \times \frac{min(\hat{t}, \hat{b})}{max(\hat{t}, \hat{b})}}. \#(33.)$$

The score $CS(i, j)$ is up to one when the distance between the object center and the corresponding position (x, y) is zero, i.e., $\hat{l} = \hat{r}$, $\hat{t} = \hat{b}$, and $\mathbb{I}(\hat{o}_{(i,j)}) = 1$. The larger the distance, the lower the score $CS(i, j)$. While the center-ness loss $\mathcal{L}_{cen}$ is defined as

$$\mathcal{L}_{cen} = \frac{-1}{\sum \mathbb{I}(\hat{o}_{(i,j)})} \sum_{\mathbb{I}(\hat{o}_{(i,j)})==1} CS(i,j) * log\, A_{cen}(i,j) + (1 - CS(i,j)) * log(1 - A_{cen}(i,j)). \#(34.)$$

4. Experimental results and analysis

4.1. Data

SENSE is trained and tested on HS datasets [19] provided by the Hyperspectral Object Tracking Competition (HOTC), which comprises 40 groups of training data and 35 groups of testing data. Each group of data consists of three types of videos with a frame rate of 25 frames per second (FPS), namely HS video, false color video, and RGB video, with 16, three, and three bands, respectively. The false color video is generated from the corresponding HS video, while the RGB video is captured from a viewpoint close to the HS video. Consequently, ground truth labels are shared between HS and false color videos, whereas labels for RGB videos are generated independently. There are 11 fine attributes including background clutter (BC), deformation (DEF), fast motion (FM), in-plane rotation (IPR), illumination variation (IV), low resolution (LR), motion blur (MB), occlusion (OCC), out-of-plane rotation (OPR), out-of-view (OV), and scale variation (SV). In addition, frame-level annotation with horizontal bounding boxes is used for evaluation.

4.2. Implementation detail

The proposed method is implemented in Python with PyTorch and trained on an RTX 4060 card. The initial Siamese

network [32] is trained with RGB datasets from YouTube-BB [68], ImageNet-VID, ImageNet-DET [37], and COCO [69]. Then, it is further trained on the HOTC HS dataset using the stochastic gradient descent optimizer with an initial learning rate of 0.001. The batch size is set to 16, and a total of 20 epochs are executed during the training process. For multi-task loss weight, γ_1, γ_2, γ_3, and γ_4 are empirically set to 1.2, 3.0, 1.0, and 1.0, respectively. While τ is set to 29.0. The training sample pairs are of sizes 320×320×16 and 128×128×16 pixels, respectively.

4.3. Assessment metric

We employ both precision and success plots to measure the performance of trackers in one-pass evaluation [70]. The precision plot displays the percentage of frames whose center location error v is less than thresholds varied from 1 to 50 pixels, and v is defined as

$$v = \sqrt{(x - X)^2 + (y - Y)^2}, \#(35. \quad)$$

where (x, y) and (X, Y) represent the center of the predicted 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

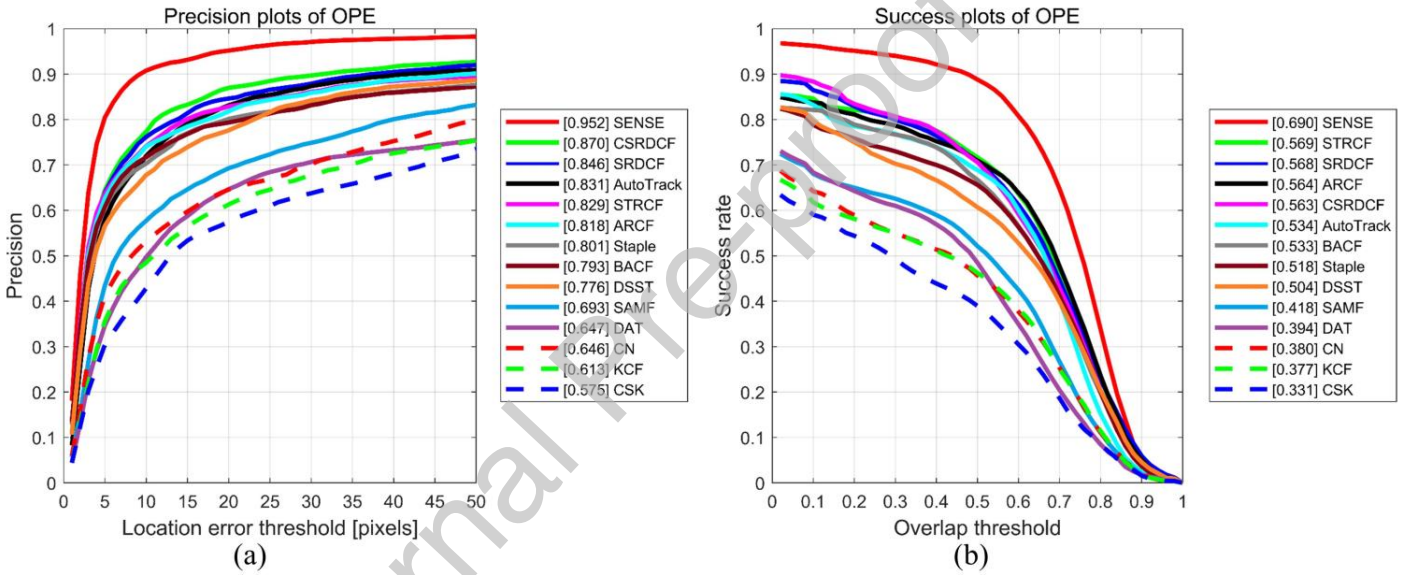


Fig. 11. Comparison with hand-crafted feature-based trackers on RGB videos. (a) Precision plot. (b) Success plot.

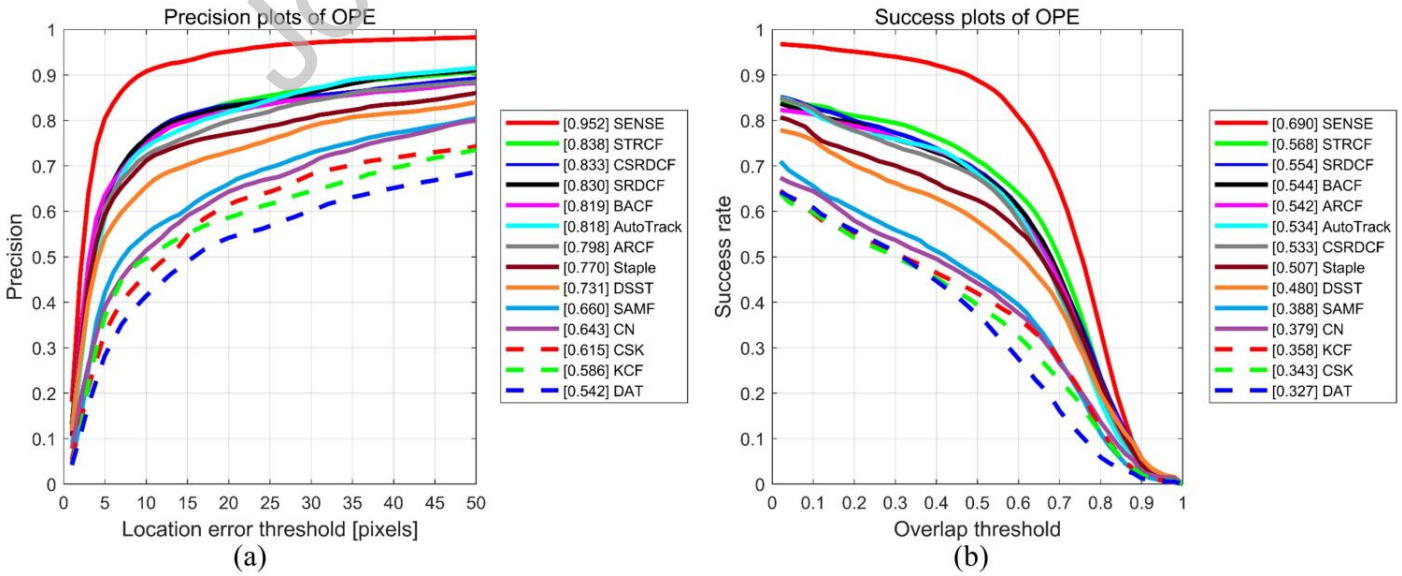


Fig. 12. Comparison with hand-crafted feature-based trackers on false color videos. (a) Precision plot. (b) Success plot.

thresholds varied from 0 to 1. Given r_t and r_g, s can be computed by

$$s = \frac{|r_t \cap r_g|}{|r_t \cup r_g|}, \#(36. \quad)$$

where ∪ and ∩ are union and intersection operators, and $|\cdot|$ stands for the number of pixels in a given region. The trackers are ranked by the precision at 20 pixels of the precision plot and the area under the curve of the success plot, i.e., Pre and Suc, respectively. While FPS is used to measure the running speed.

4.4. Comparison with RGB trackers

4.4.1. Hand-crafted feature-based trackers

We compare the SENSE with 13 hand-crafted feature-based SOTAs including CSK [71], CN [72], SAMF [73], DAT [74], KCF [4], SRDCF [75], Staple [76], DSST [77], BACF [78], CSRDCF [79], STRCF [80], ARCF [81], and AutoTrack [82]. SENSE is evaluated on HS videos, while the others are tested on RGB videos and false color videos. Table 1 details t h e

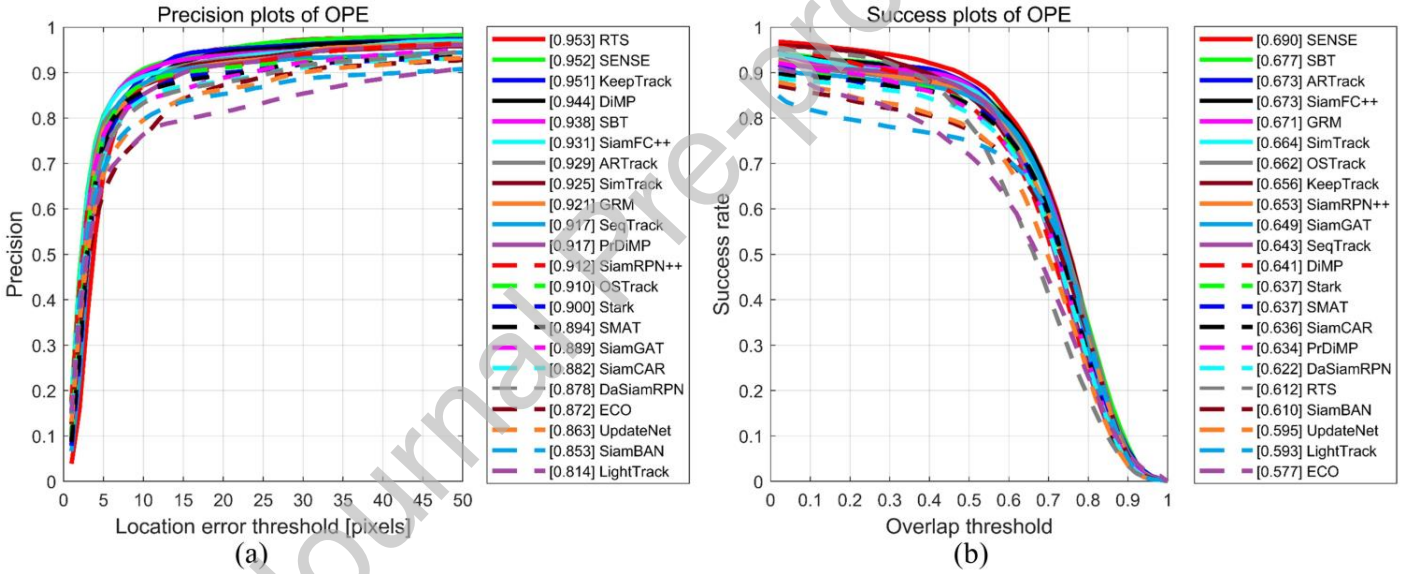


Fig. 13. Comparison with deep feature-based trackers on RGB videos. (a) Precision plot. (b) Success plot.

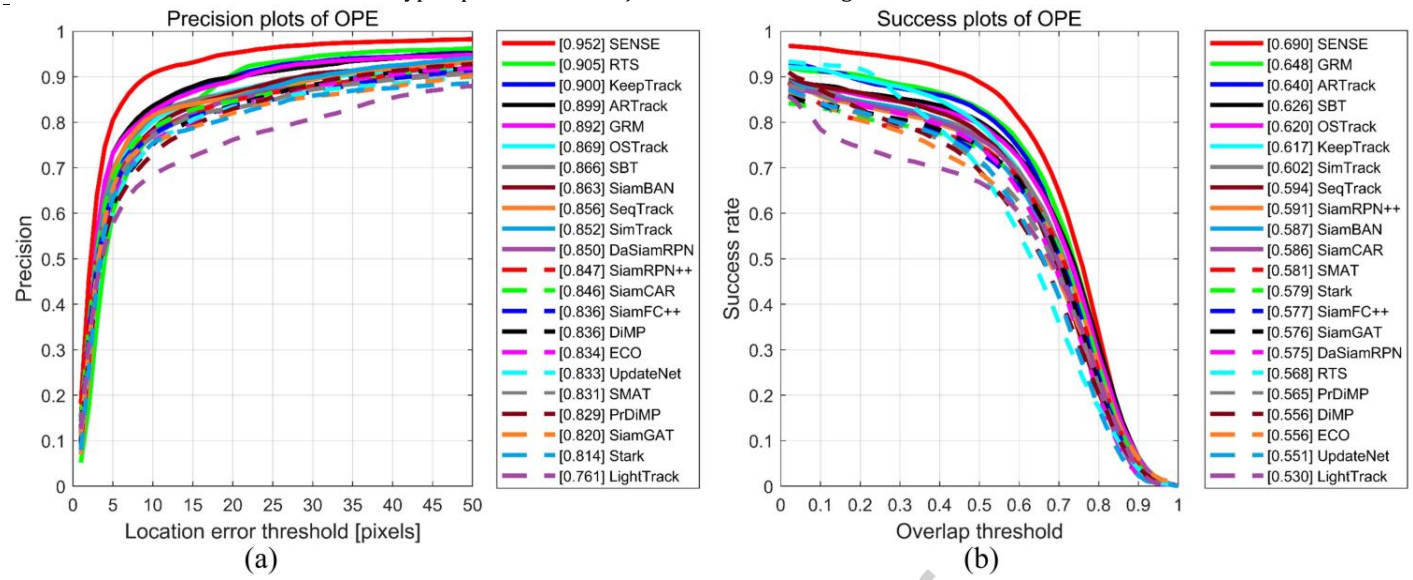


Fig. 14. Comparison with deep feature-based trackers on false color videos. (a) Precision plot. (b) Success plot.

results and characteristics of trackers. Fig. 11 presents the precision and success plots tested on RGB videos. As evident in Table 1 and Fig. 11, SENSE obtains the optimal results with a Pre of 0.952 and a Suc of 0.690. Compared to STRCF, SENSE exhibits improvements in Pre and Suc by 12.3% and 12.1%, respectively. Compared to SRDCF and ARCF, SENSE achieves impressive improvements in Suc of 12.2% and 12.6%, respectively. The results highlight the potential of leveraging material and motion cues in synergistic ways. It is conceivable to adapt the RGB tracker to HS video by converting the HS video into a false color video. Naturally, we conduct experiments on the false color video, and Fig. 12 and Table 1 report the results. STRCF maintains a respectable performance followed by SRDCF with Suc of 0.568 and 0.554, respectively. In contrast, SENSE achieves gains of 12.2% and 13.6%, respectively. This is attributed to the adaptive acquisition of complementary false modalities with varying contributions facilitated by the SSSE module, which are then aggregated and enhanced by the CFMF module. Hence, SENSE adeptly harnesses the abundant spectral information. Furthermore, the incorporation of an MA module equips SENSE to overcome challenges associated with unreliable material cues during abnormal states. Notably, the ranking of the compared trackers remains largely consistent with that of the RGB video, besides yielding lower results. This discrepancy can be attributed to inherent different spectral characteristics between false color and RGB videos, despite

Table 1

Parallel comparison with RGB trackers listed in chronological order. The top three scores are marked in red, green, and blue, respectively.

Tracker	Venue	Feature/Backbone	RGB		FAC/HS		PreD	SucD
			Pre	Suc	Pre	Suc		
CSK [71]	ECCV 2012	I	0.575	0.331	0.615	0.343	-4.0%	-1.2%
CN [72]	CVPR 2014	CN+I	0.646	0.380	0.643	0.379	0.3%	0.1%
SAMF [73]	ECCV 2015	HOG+CN+I	0.693	0.418	0.660	0.388	3.3%	**3.0%**
DAT [74]	CVPR 2015	CH	0.647	0.394	0.542	0.327	**10.5%**	**6.7%**
KCF [4]	TPAMI 2015	HOG	0.613	0.377	0.586	0.358	2.7%	1.9%
SRDCF [75]	ICCV 2015	HOG	**0.846**	**0.568**	**0.830**	**0.554**	1.6%	1.4%
Staple [76]	CVPR 2016	HOG+CN	0.801	0.518	0.770	0.507	3.1%	1.1%
DSST [77]	TPAMI 2017	HOG+I	0.776	0.504	0.731	0.480	**4.5%**	2.4%

BACF [78]	ICCV 2017	HOG	0.793	0.533	0.819	**0.544**	-2.6%	-1.1%
CSRDCF [79]	IJCV 2018	HOG+CN+CH	**0.870**	0.563	**0.833**	0.533	**3.7%**	**3.0%**
STRCF [80]	CVPR 2018	HOG+CN	0.829	**0.569**	**0.838**	**0.568**	-0.9%	0.1%
ARCF [81]	ICCV 2019	HOG+CN+I	0.818	**0.564**	0.798	0.542	2.0%	2.2%
AutoTrack [82]	CVPR 2020	HOG+CN+I	**0.831**	0.534	0.818	0.534	1.3%	0.0%
ECO [83]	CVPR 2017	VGG-M	0.872	0.577	0.834	0.556	3.8%	2.1%
DaSiamRPN [84]	ECCV 2018	AlexNet	0.878	0.622	0.850	0.575	2.8%	4.7%
DiMP [85]	ICCV 2019	ResNet-50	**0.944**	0.641	0.836	0.556	**10.8%**	**8.5%**
SiamRPN++ [86]	CVPR 2019	ResNet-50	0.912	0.653	0.847	0.591	6.5%	6.2%
UpdateNet [87]	ICCV 2019	AlexNet	0.863	0.595	0.833	0.551	3.0%	4.4%
PrDiMP [88]	CVPR 2020	ResNet-50	0.917	0.634	0.829	0.565	**8.8%**	6.9%
SiamBAN [89]	CVPR 2020	ResNet-50	0.853	0.610	0.863	0.587	-1.0%	2.3%
SiamFC++ [90]	AAAI 2020	AlexNet	0.931	**0.673**	0.836	0.577	**9.5%**	**9.6%**
KeepTrack [91]	ICCV 2021	ResNet-50	**0.951**	0.656	**0.900**	0.617	5.1%	3.9%
SiamGAT [92]	CVPR 2021	GoogLeNet	0.889	0.649	0.820	0.576	6.9%	**7.3%**
LightTrack [93]	CVPR 2021	Custom	0.814	0.593	0.761	0.530	5.3%	6.3%
Stark [94]	ICCV 2021	ResNet-50	0.900	0.637	0.814	0.579	8.6%	5.8%
RTS [95]	ECCV 2022	ResNet-50	**0.953**	0.612	**0.905**	0.568	4.8%	4.4%
SiamCAR [32]	IJCV 2022	ResNet-50	0.882	0.636	0.846	0.586	3.6%	5.0%
OSTrack [34]	ECCV 2022	ViT-Base	0.910	0.662	0.869	0.620	4.1%	4.2%
SimTrack [96]	ECCV 2022	ViT-Base	0.925	0.664	0.852	0.602	7.3%	6.2%
SBT [97]	CVPR 2022	SBT-Base	0.938	**0.677**	0.866	**0.626**	7.2%	5.1%
GRM [98]	CVPR 2023	ViT-Large	0.921	**0.671**	0.892	**0.648**	2.9%	2.3%
SeqTrack [99]	CVPR2023	ViT-Base	0.917	0.643	0.856	0.594	6.1%	4.9%
ARTrack [56]	CVPR2023	ViT-Base	0.929	**0.673**	**0.899**	**0.640**	3.0%	3.3%
SMAT [100]	WACV 2024	MobileViTv2	0.894	0.637	0.831	0.581	6.3%	5.6%
SENSE	Ours	ResNet-50	n/a	n/a	**0.952**	**0.690**	n/a	n/a

RGB, FAC, and HS represent the red-green-blue, false color, and hyperspectral videos, respectively. PreD and SucD denote 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 the dashed line. SENSE is shown at the bottom. n/a stands for not applicable.

sharing three bands. Experimental results also underscore the suboptimal nature of converting HS video into false color video, as it unavoidably results in loss and distortion of crucial material information that is pivotal for achieving robust performance.

4.4.2. Deep feature-based trackers

In contrast to hand-crafted features, deep features are more discriminative and have achieved considerable advances in the RGB tracking domain. Here we compare SENSE with 21 deep feature-based SOTAs including ECO [83], DaSiamRPN [84], DiMP [85], SiamRPN++ [86], UpdateNet [87], PrDiMP [88], SiamBAN [89], SiamFC++ [90], KeepTrack [91], SiamGAT [92], LightTrack [93], Stark [94], RTS [95], SiamCAR [32], OSTrack [34], SimTrack [96], SBT [97], GRM [98], SeqTrack [99], ARTrack [56], and SMAT [100], covering a variety of backbones and tracking paradigms.

Table 1 reports results tested on RGB and false color videos. Precision and success plots are shown in Fig. 13 and Fig. 14. These trackers are tailored for RGB video, therefore, exhibit superior performance in RGB compared to false-color videos, and a similar conclusion can be drawn from Fig. 15. However, deep feature-based trackers demonstrate a more

pronounced decline in Suc when transitioning from RGB to false color videos compared to trackers that rely on hand-crafted features, as shown in Fig. 15. This can be attributed to the significant disparities between RGB and false color videos (Fig. 16), with the performance of deep trackers highly dependent on the specific characteristics of the training data, which is typically in RGB format. Furthermore, it is observed that trackers utilizing deep features consistently outperform those utilizing hand-crafted features. This can be attributed to the capability of complex networks to learn discriminative and generalized object representations from massive RGB data. SENSE can learn the rich material cue present in HS data and integrate the motion cue, achieving competitive results.

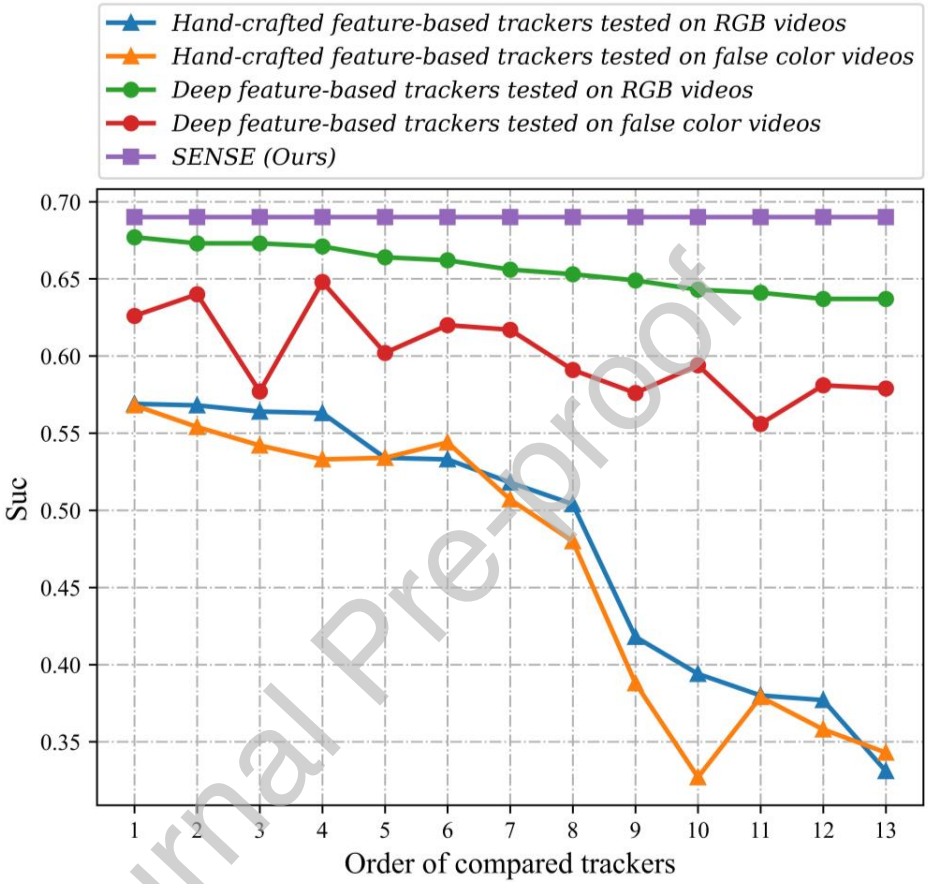


Fig. 15. Parallel comparisons with RGB trackers using hand-crafted and deep features. These trackers are ranked by the Suc tested on RGB videos. The hand-crafted feature-based trackers are STRCF, SRDCF, ARCF, CSRDCF, AutoTrack, BACF, Staple, DSST, SAMF, DAT, CN, KCF, and CSK, respectively. The deep feature-based trackers are SBT, ARTrack, SiamFC++, GRM, SimTrack, OSTrack, KeepTrack, SiamRPN++, SiamGAT, SeqTrack, DiMP, SMAT, and Stark, respectively.

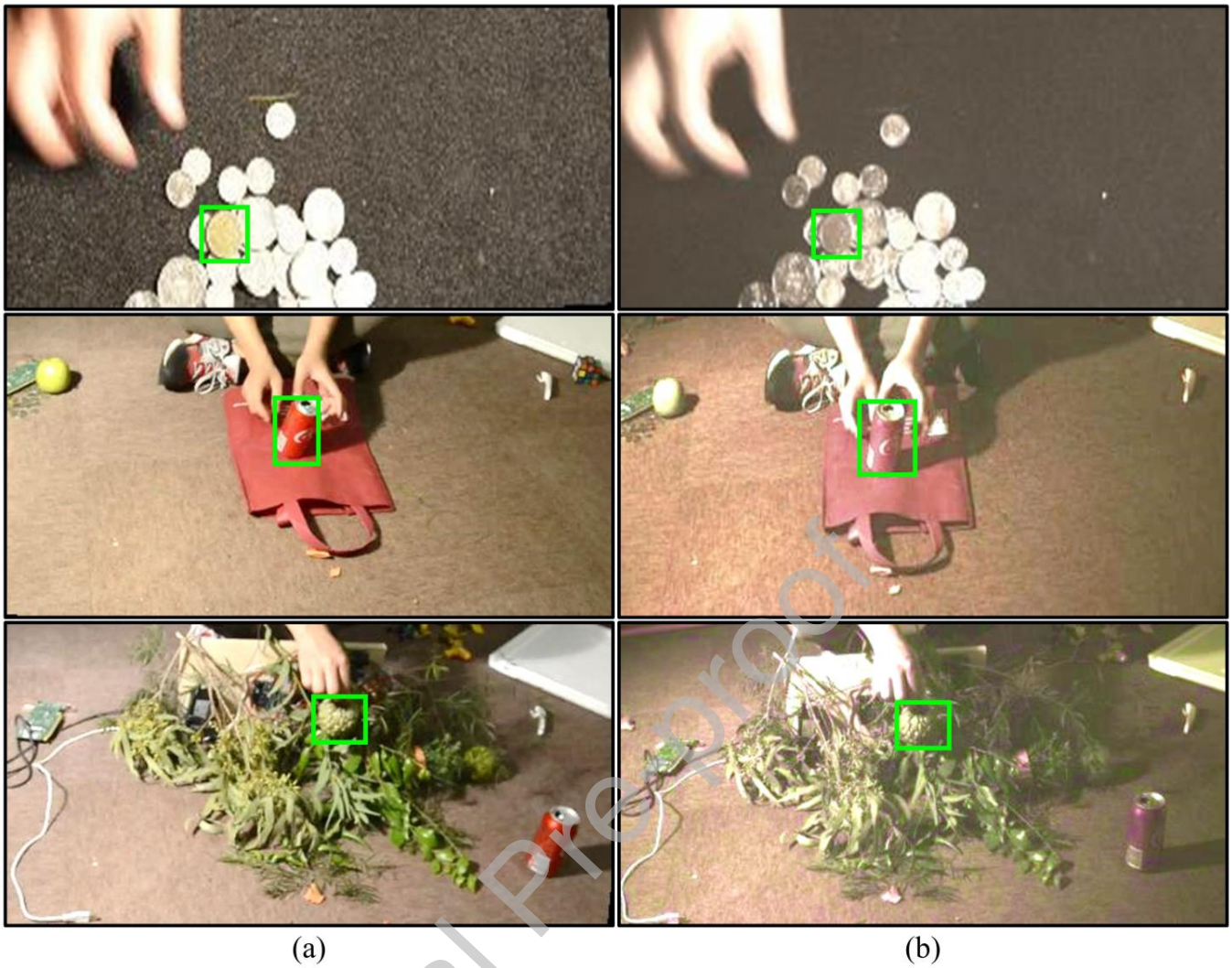


Fig. 16. Visualization of RGB data and false color data generated from HS data. (a) RGB data. (b) False color data.

4.5. Comparison with hyperspectral trackers

Further, we compare SENSE with 16 SOTA HS trackers including CNHT [60], DeepHKCF [101], MHT [19], BAE-Net [22], MFI [59], SST-Net [28], SiamHYPER [8], TSCFW [24], BAHT [38], TASSCF [25], DeepTASSCF [25], SEE-Net [21], SiamOHOT [31], SiamBAG [23], SiamHT [40], and SPIRIT [42], covering Siamese network, VITIAL, and correlation filter paradigms. Experimental results and characteristics of trackers are summarized in Table 2. Fig. 17 shows the precision and success plots. For the Pre, SENSE, SiamHYPER, SEE-Net, and SPIRIT rank among the top four with scores of 0.952, 0.947, 0.934, and 0.925, respectively. For the Suc, SPIRIT, SiamHYPER, SEE-Net, and BAHT produce competitive outcomes with scores of 0.679, 0.678, 0.666, and 0.665, respectively, securing the top four positions among compared trackers. While SENSE shows satisfactory results surpasses them by 1.1%, 1.2%, 2.4%, and 2.5%, respectively. Overall, SENSE achieves optimal performance in both Pre and Suc, underscoring the potential of integrating material and motion cues in HS video object tracking.

Moreover, Table 2 highlights that the top-ranked trackers usually inherit the Siamese network, while the low-ranked trackers adopt the correlation filter, indicating that the Siamese network may be superior to the correlation filter in HS video object tracking. We can draw a similar conclusion in the evolution of the RGB trackers from Table 1. In particular, recent works such as SPIRIT, SiamBAG, SEE-Net, and SiamHYPER strive to leverage the rich material cue for robust object modeling. The ideal has been validated because the abundance of spectral information potentially enhances the material awareness of trackers. In our approach, the SSSE module enables SENSE to adaptively utilize spectral

information by

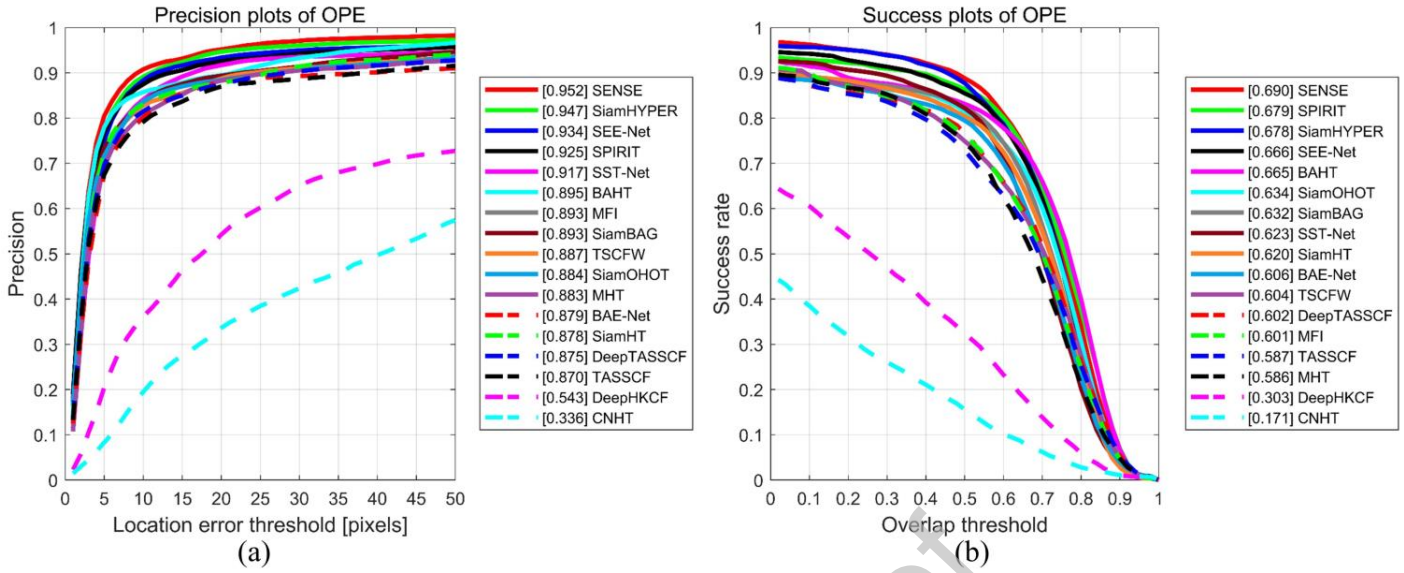


Fig. 17. Comparison with hyperspectral trackers on hyperspectral videos. (a) Precision plot. (b) Success plot.

Table 2

Characteristics and results of HS trackers.

Tracker	Venue	Framework	Feature	UFB	UMC	Pre	Suc	FPS	MOP
CNHT [60]	ICSM 2018	KCF	Deep feature	✓	-	0.336	0.171	2.6	CPU
DeepHKCF [101]	TGRS 2019	KCF	Deep feature	-	-	0.543	0.303	0.9	CPU
MHT [19]	TIP 2019	KCF	Hand-crafted feature	✓	-	0.883	0.586	2.2	CPU
BAE-Net [22]	ICIP 2020	VITAL	Deep feature	✓	-	0.879	0.606	0.5	GPU
MFI [59]	WISP 2021	KCF	Deep + Hand-crafted features	-	-	0.893	0.601	0.4	CPU
SST-Net [28]	WISP 2021	VITAL	Deep feature	✓	-	0.917	0.623	0.5	GPU
SiamHYPER [8]	TIP 2022	SiamFC	Deep feature	✓	-	**0.947**	**0.678**	**19.0**	GPU
TSCFW [24]	TGRS 2022	KCF	Hand-crafted features	✓	-	0.887	0.604	3.4	CPU
BAHT [38]	GRSL 2022	SiamFC	Deep feature	-	-	0.895	0.665	16.0	GPU
TASSCF [25]	CVIU 2022	KCF	Hand-crafted features	-	-	0.870	0.587	16.0	CPU
DeepTASSCF [25]	CVIU 2022	KCF	Deep feature	-	-	0.875	0.602	6.0	CPU
SEE-Net [21]	TIP 2023	SiamFC	Deep feature	✓	-	**0.934**	0.666	8.7	GPU
SiamOHOT [31]	TGRS 2023	SiamFC	Deep feature	✓	-	0.884	0.634	**38.0**	GPU
SiamBAG [23]	TGRS 2023	SiamFC	Deep feature	✓	-	0.893	0.632	5.7	GPU
SiamHT [40]	NCA 2023	SiamFC	Deep feature	-	-	0.878	0.620	16.0	GPU
SPIRIT [42]	TGRS 2024	SiamFC	Deep feature	✓	-	0.925	**0.679**	**26.0**	GPU
SENSE	Ours	SiamFC	Deep feature	✓	✓	**0.952**	**0.690**	15.4	GPU

KCF denotes the kernelized correlation filter [4], VITAL stands for visual tracking via adversarial learning [102], and SiamFC represents the fully convolutional Siamese network [5]. UFB denotes the attempt to use of full band. UMC stands for the use of motion cues. MOP is the main operation platform. WISP denotes the WHISPERS.

generating complementary false modalities with varying contributions. Furthermore, we propose a CFMF module to

aggregate and enhance the features learned from false modalities. It is worth noting that current HS trackers often ignore the motion cue contained in sequential frames, which can be particularly effective when the material cue is unreliable such as occlusion. With this in mind, we propose an MA module to enable SENSE to sense and handle abnormal states, arousing motion awareness.

In general, SENSE builds upon the Siamese network and fuses the material and motion cues to learn robust object representations while modeling object motion and shape, culminating in material and motion awareness capabilities. In addition, SENSE can maintain a competitive speed of 15.4 FPS. In particular, SENSE, SiamOHOT, and SEE-Net all inherit SiamFC, but with significant differences in running speed. This is because SiamOHOT places significant emphasis on employing knowledge distillation to refine the model in order to enhance its efficiency. SEE-Net uses decision-level fusion to improve effectiveness but results in an expensive computational burden. While SENSE incorporates the feature-level fusion to improve efficiency, and designs the spectral-spatial self-expression module, cross-false modality fusion module, and motion awareness module for improving effectiveness.

4.6. Attribute-based evaluation

Here, we further perform attribute-based evaluation with 13 HS trackers including CNHT, DeepHKCF, MHT, BAE-Net, MFI, SST-Net, SiamHYPER, TSCFW, TASSCF, DeepTASSCF, SEE-Net, SiamOHOT, and SiamBAG and eight RGB trackers including CSRDCF, STRCF, ARCF, AutoTrack, GRM, SeqTrack, ARTrack, and SMAT. The RGB trackers are tested on false color videos generated from HS videos.

Table 3 and Table 4 show the Pre and Suc scores, and Fig. 18 and Fig. 19 present the precision and success plots, respectively. For the Pre, SENSE obtains a top three for seven (i.e., BC, FM, IV, LR, MB, OCC, and SV) out of 11 attributes and achieves the first overall (i.e., OVE). For the Suc, SENSE obtains a top three for nine (i.e., BC, FM, IPR, IV, LR, MB, OCC, OPR, and SV) out of 11 attributes. OCC is a challenging attribute that is difficult to be addressed through the material cue alone, particularly in full occlusion scenario with unreliable material cues. Benefiting from the MA module, SENSE can

Table 3

Results of per-attribute and overall in terms of the Pre metric.

Tracker	Venue	BC	DEF	FM	IPR	IV	LR	MB	OCC	OPR	OV	SV	OVE
CSRDCF [79]	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
STRCF [80]	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
ARCF [81]	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 [82]	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
GRM [98]	CVPR 2023	0.873	0.945	0.979	0.974	0.860	0.802	0.966	**0.899**	**0.979**	**1.000**	0.895	0.892
SeqTrack [99]	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
ARTrack [56]	CVPR2023	0.875	0.945	0.982	**0.985**	0.846	0.888	0.976	0.868	0.962	**1.000**	**0.920**	0.899
SMAT [100]	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 [60]	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 [101]	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 [19]	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
BAE-Net [22]	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
MFI [59]	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
SST-Net [28]	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
SiamHYPER [8]	TIP 2022	**0.965**	**0.956**	**0.993**	0.943	**0.916**	**0.983**	**0.999**	**0.909**	0.943	0.891	**0.920**	**0.947**
TSCFW [24]	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 [25]	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 [25]	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
SEE-Net [21]	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**

SiamOHOT [31]	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 [23]	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
SENSE	Ours	**0.966**	0.948	**0.992**	0.976	**0.922**	**0.946**	**0.993**	**0.921**	0.971	0.891	**0.946**	**0.952**

The OVE indicates the overall score.

Table 4

Results of per-attribute and overall in terms of the Suc metric.

Tracker	Venue	BC	DEF	FM	IPR	IV	LR	MB	OCC	OPR	OV	SV	OVE
CSRDCF [79]	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
STRCF [80]	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
ARCF [81]	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 [82]	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
GRM [98]	CVPR 2023	0.640	**0.732**	0.674	**0.742**	**0.612**	0.563	0.675	**0.645**	**0.749**	**0.837**	**0.653**	0.648
SeqTrack [99]	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
ARTrack [56]	CVPR2023	0.629	**0.721**	**0.675**	**0.751**	0.573	0.600	0.665	0.621	**0.730**	**0.841**	**0.651**	0.640
SMAT [100]	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 [60]	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 [101]	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 [19]	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
BAE-Net [22]	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
MFI [59]	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
SST-Net [28]	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
SiamHYPER [8]	TIP 2022	**0.714**	**0.721**	**0.711**	0.721	**0.586**	**0.664**	**0.753**	**0.634**	0.714	0.602	0.646	**0.678**
TSCFW [24]	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 [25]	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 [25]	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
SEE-Net [21]	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**
SiamOHOT [31]	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 [23]	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
SENSE	Ours	**0.721**	0.708	**0.746**	**0.752**	**0.609**	**0.622**	**0.730**	**0.645**	**0.737**	0.629	**0.674**	**0.690**

sense the abnormal state and exploit the motion cue to address it. As a result, SENSE obtains the optimal Pre and Suc of 0.921 and 0.645, which are 3.3% and 2.3% higher than SEE-Net. Moreover, BC is another demanding attribute where the background has a similar texture or color as the object. In this attribute, HS trackers such as SENSE, SEE-Net, and MFI significantly outperform RGB trackers such as GRM, SeqTrack, and ARCF. This is attributed to the capability of HS trackers to uncover the rich material cue to discriminate the physical material of objects. In particular, SENSE achieves the highest Suc of 0.721 higher than that of SEE-Net and MFI by 1.6% and 7.0%, respectively, thanks to strong material awareness and additional motion awareness capabilities benefitted from the proposed SSSE, CFMF, and MA modules.

Overall, SENSE can combine the material cue with the motion cue to achieve superior performance under challenging attributes, validating the importance of jointly considering material awareness and motion awareness in the HS tracking domain.

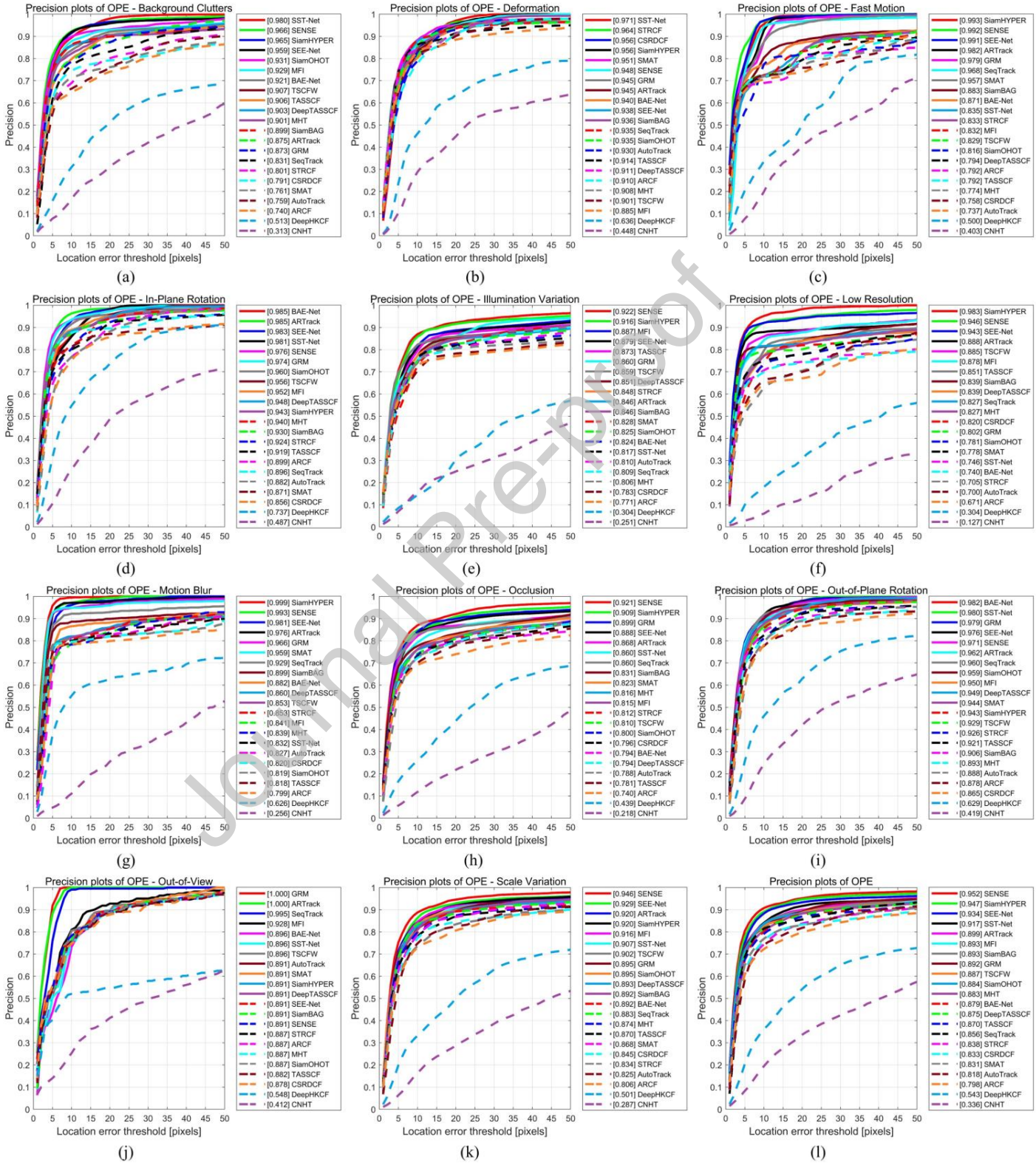


Fig. 18. The precision plot of each attribute and overall. (a) BC. (b) DEF. (c) FM. (d) IPR. (e) IV. (f) LR. (g) MB. (h) OCC. (i) OPR. (j) OV. (k) SV. (l) OVE.

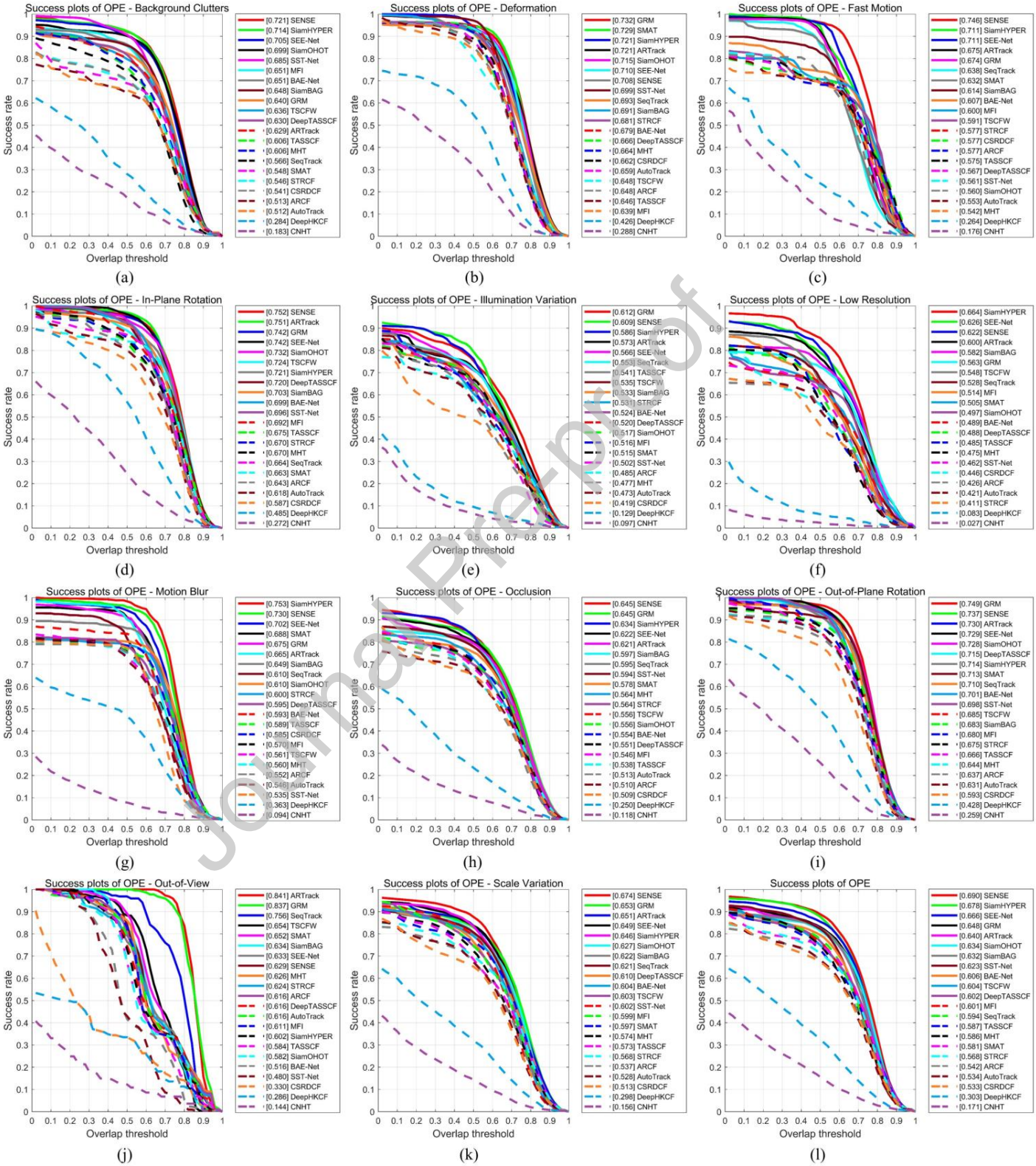


Fig. 19. The success plot of each attribute and overall. (a) BC. (b) DEF. (c) FM. (d) IPR. (e) IV. (f) LR. (g) MB. (h) OCC. (i) OPR. (j) OV. (k) SV. (l) OVE.

4.7. Visual comparison

We conduct qualitative comparisons with HS trackers including SPIRIT, SiamHYPER, SiamOHOT, DeepHKCF, MHT, and CNHT and RGB trackers including STRCF and GRM. Fig. 20 shows qualitative results. In the *bus2* video, an object

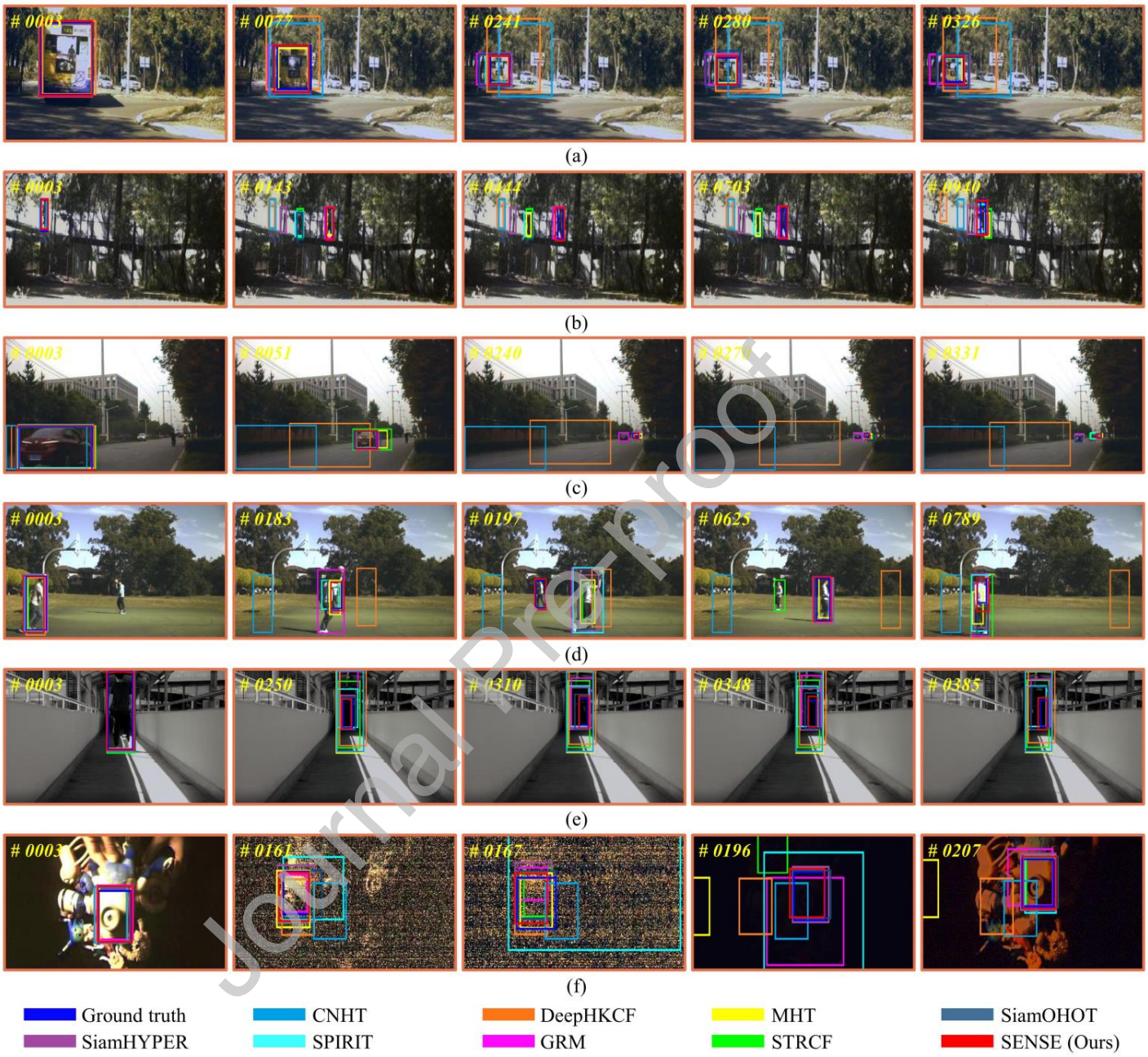


Fig. 20. Qualitative comparisons with SOTA trackers. Results are displayed in corresponding false color images with the current frame shown in the upper-left corner. (a) *bus2*, attribute: OCC, SV, FM, and IV. (b) *campus*, attribute: OCC, IV, and SV. (c) *car3*, attribute: OCC, SV, IV, and LR. (d) *playground*, attribute: OCC and SV. (e) *student*, attribute: SV and IV. (f) *toy2*, attribute: OCC, SV, OPR, BC, and IV.

encounters the OCC, SV, FM, and IV. CNHT and DeepHKCF lose track of the object initially. SiamOHOT, SiamHYPER, and GRM struggle to accurately estimate the scale. SENSE demonstrates superior performance in locating the object and estimating its scale, due to its capability of adaptively generating, aggregating, and enhancing complementary false modalities via the SSSE and CFMF modules. In the *playground* video, the occlusion and scale variation problems are significant, and the tracked object is surrounded by a similar object, which challenges the tracking algorithms. In this

scenario, most trackers such as SPIRIT, MHT, GRM, and STRCF lose the object to varying degrees. Benefitting from the MA module, SENSE can effectively mitigate these interferences by fusing the material and motion cues, hence successfully locating the object, as shown in frame #0197. In the other cases shown in Fig. 20, SENSE also exhibits remarkable performance. The results highlight its robustness and accuracy, rendering it an ideal candidate for HS object tracking.

5. Ablation study

The major contributions of SENSE include SSSE, CFMF, and MA modules. The SSSE module and CFME module focus on mining the physical material cues of the object, while the MA module emphasizes the motion cues. Here we will discuss their effects. To this end, we construct ten models including Model-0, Model-1, Model-2, Model-3, Model-4, Model-5, Model-6, Model-7, Model-8, and Model-9. Table 5 details the components and results of different models. Fig. 21 shows the precision and success plots. Without consideration of rich material and motion cues, the baseline Model-0 [32] is performed on HS videos by converting them into false color videos, while the remaining models are tested on HS videos. Comprehensive analysis and discussion are shown as follows.

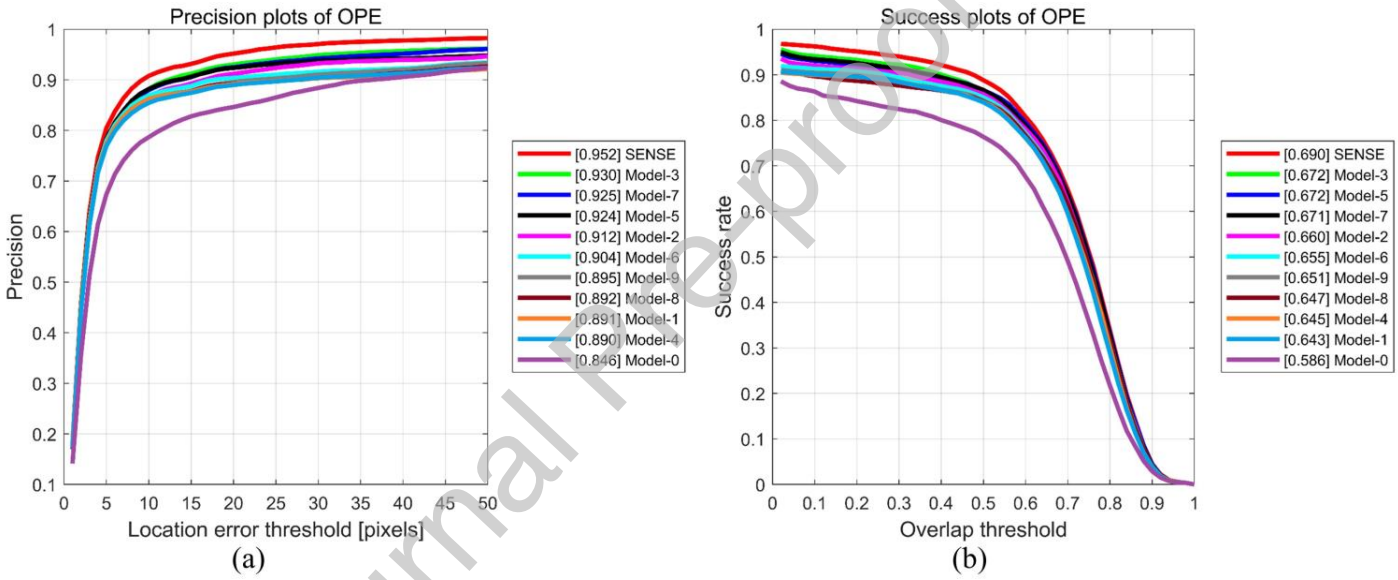


Fig. 21. Ablation result of different models. (a) Precision plot. (b) Success plot.

Table 5

Composition and results of different models.

Model	SSSE	CFMF			MA	Pre	Suc	PreI	SucI
		IC	DF	CF					
Model-0	-	-	-	-	-	0.846	0.586	n/a	n/a
Model-1	✓	✓	-	-	-	0.891	0.643	4.5%	5.7%
Model-2	✓	✓	-	-	✓	0.912	0.660	2.1%	7.4%
Model-3	✓	-	✓	✓	✓	**0.930**	**0.672**	**8.4%**	**8.6%**
Model-4	✓	-	✓	✓	-	0.890	0.645	4.4%	5.9%
Model-5	✓	✓	✓	-	✓	0.924	**0.672**	7.8%	**8.6%**
Model-6	✓	✓	✓	-	-	0.904	0.655	5.8%	6.9%
Model-7	✓	✓	-	✓	✓	**0.925**	**0.671**	**7.9%**	**8.5%**
Model-8	✓	✓	-	✓	-	0.892	0.647	4.6%	6.1%

Model-9	✓	✓	✓	✓	-	0.895	0.651	4.9%	6.5%
SENSE	✓	✓	✓	✓	✓	**0.952**	**0.690**	**10.6%**	**10.4%**

SSSE, CFMF, and MA are the proposed spectral-spatial self-expression module, cross-false modality fusion module, and motion awareness module, respectively. IC, DF, and CF of the CFMF module are the use of the initial contributions, differential features, and common features, respectively. PreI and SucI are the Pre and Suc improvements of the current model compared to Model-0, respectively.

5.1. Effectiveness of SSSE module

Based on Model-0, Model-1 incorporates the SSSE module that can adaptively divide the HS image into complementary false modalities according to their contributions. The contributions naturally can be seen as initial fusion weights of false modalities.

From Table 5, it can be observed that Model-0 yields the lowest Pre of 0.846 and Suc of 0.586. With the introduction of the SSSE module and initial contributions (IC) in Model-1, there are significant gains in both Pre and Suc by 4.5% (from 0.846 to 0.891) and 5.7% (from 0.586 to 0.643), respectively. It can be attributed to the adaptive grouping capability of SSSE and the fusion of IC.

5.2. Effectiveness of CFMF module

In this section, we conduct experiments to validate the CFMF module. It initially merges the false modalities based on the contributions generated by the SSSE module and subsequently aggregates and enhances the differential-common features for cross-false modality fusion.

For the IC, a comparison between Model-9 and Model-4 reveals that using the IC improves Pre and Suc by 0.5% and 0.6%, respectively. Similarly, comparing SENSE and Model-3 shows gains of 2.2% in Pre and 1.8% in Suc with the addition of IC. For the differential features (DF), we compare Model-5 and Model-2 and find that the addition of DF results in a 1.2% improvement in both Pre and Suc. Conversely, removing DF in Model-1 leads to a reduction of 1.3% in Pre and 1.2% in Suc when compared to Model-6. Comparisons in Model-9 & Model-8 and SENSE & Model-7 further confirm the importance of DF. Additionally, the introduction of the common features (CF) enhances the Pre and Suc of Model-7 by 1.3% and 1.1%, respectively, compared to Model-2. Notably, Model-5 exhibits a lower Pre of 0.924 and Suc of 0.672, indicating a 2.8% and 1.8% decrease compared to SENSE due to the absence of CF. Similar conclusions can be drawn from the comparison between Model-1 and Model-8.

To comprehensively validate the CFMF module, we compare Model-9 and Model-1, where Pre and Suc are boosted by 0.4% and 0.8%, respectively. Compared to Model-2, SENSE demonstrates significant improvements of 4.0% in Pre and 3.0% in Suc with the aid of the CFME module. In addition, the CFMF module, building upon the SSSE module, further enhances material awareness, resulting in remarkable improvements of 4.9% in Pre and 6.5% in Suc when comparing Model-0 and Model-9.

5.3. Effectiveness of MA module

MA module enables HS video object tracking from a perspective of cooperating with material and motion cues. A comparison between Model-2 and Model-1 reveals enhancements of 2.1% and 1.7% in Pre and Suc due to the inclusion of the MA module. Conversely, when comparing Model-3 and Model-4, a reduction of 4.0% and 2.7% in Pre and Suc can be observed with the removal of the MA module. Furthermore, in comparison to Model-9, SENSE obtains substantial improvements of 5.7% and 3.9% in Pre and Suc, respectively. Similar trends are evident in comparisons of Model-7 & Model-8 and Model-5 & Model-6, which affirm the effectiveness of the MA module.

In addition, it is observed that the role of the MA module is prominent based on SSSE and CFMF modules, which further validates the potential of synergizing material and motion cues in the HS video tracking domain. Specifically, SENSE achieves Pre and Suc of 0.952 and 0.690, higher than that of Model-0 by 10.6% and 10.4%, respectively. Overall,

extensive experiments have witnessed the effects of SSSE, CFMF, and MA modules, proving the competitive performance of SENSE.

5.4. Adaptive selection of gamma

In this section, we explain the current parameter selection and further conduct an adaptive gamma selection experiment, and compare it with the current method. In SENSE, γ_2, γ_3, and γ_4 are set with the same weights as the baseline [32]. γ_3 is empirically set to 1.2 to favour the learning of modality generation with the consideration of classification and regression

Table 6

Comparison with an adaptive selection of gamma weights on overall and per-attribute.

Model	OVE		BC		DEF		FM		IPR		IV	
	Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc	Pre	Suc
AGS	0.940	0.674	0.973	0.717	0.947	0.706	0.995	0.734	0.977	0.752	0.871	0.561
SENSE	0.952	0.690	0.966	0.721	0.948	0.708	0.992	0.746	0.976	0.752	0.922	0.609
Model	**LR**		**MB**		**OCC**		**OPR**		**OV**		**SV**	
	Pre	**Suc**	**Pre**	**Suc**	**Pre**	**Suc**	**Pre**	**Suc**	**Pre**	**Suc**	**Pre**	**Suc**
AGS	0.947	0.603	0.993	0.731	0.900	0.625	0.971	0.736	0.891	0.570	0.923	0.652
SENSE	0.946	0.622	0.993	0.730	0.921	0.645	0.971	0.737	0.891	0.629	0.946	0.674

capabilities. Further, we conduct experiments and compare with the adaptive gamma selection method (AGS). As shown in Table 6, we re-train the network to have the gamma weights being adaptively chosen by dividing each loss by the average loss. The fashion of choosing gamma is not always effective, despite demonstrating good performance in the proposed method.

6. Conclusions

This article presents an end-to-end hyperspectral video object tracker via fusing material and motion cues (SENSE) for HS video object tracking. First, we propose an SSSE module, which captures both spectral and spatial features to efficiently solve the self-expression model and bridge the band gap. With guidance from the SSSE module, the HS image is adaptively grouped into complementary false modalities with varying contributions. These false modalities are then fed to a feature extraction module, which is pre-trained with RGB data, to address the issue of limited training samples. Additionally, a CFMF module is proposed to aggregate and enhance the differential-common features extracted from false modalities, resulting in robust object representations. Finally, we design an MA module, which includes an awareness selector to determine the reliability of the material and motion cues, as well as a motion prediction scheme to handle abnormal states such as occlusion and fast motion. By considering both material and motion cues, SENSE can arouse the material and motion awareness. Comprehensive comparisons and in-depth analysis are performed to demonstrate the superiority of our approach.

Nevertheless, the current motion model is relatively simplistic and may not fully capture complex motion patterns. Future work will focus on uncovering generalized motion cues to further improve performance. Furthermore, joining appearance space and motion state space for sequence-level training may be a promising direction for robust hyperspectral object tracking.

Acknowledgments

Sincerely thanks to the editors and reviewers for their valuable comments and constructive suggestions. This work was supported in part by the National Natural Science Foundation of China (42230108 and 42271411).

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] S. Du and S. Wang, "An Overview of Correlation-Filter-Based Object Tracking," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 1, pp. 18-31, Feb 2022.

[3] S. Liu, S. Huang, S. Wang, K. Muhammad, P. Bellavista, and J. Del Ser, "Visual tracking in complex scenes: A location fusion mechanism based on the combination of multiple visual cognition flows," *Inf. Fusion*, Article vol. 96, pp. 281-296, Aug 2023.

[4] 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.

[5] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. S. Torr, "Fully-Convolutional Siamese Networks for Object Tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV) /IEEE Trans. Signal Process.*, vol. 9914, 2016, pp. 850-865.

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

[7] Y. Chen., Y. Tang., Z. Yin., T. Han., B. Zou., and H. Feng., "Single Object Tracking in Satellite Videos: A Correlation Filter-Based Dual-Flow Tracker," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 6687-6698, 2022.

[8] 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.

[9] X. Lan, X. Gu, and X. Gu, "MMNet: Multi-modal multi-stage network for RGB-T image semantic segmentation," *Applied Intelligence*, Article vol. 52, no. 5, pp. 5817-5829, Mar 2022.

[10] S. Yan *et al.*, "DepthTrack: Unveiling the Power of RGBD Tracking," in *18th IEEE/CVF International Conference on Computer Vision (ICCV)*, Electr Network, 2021, 2021, pp. 10705-10713.

[11] 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.

[12] 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.

[13] Z. Tang, T. Xu, H. Li, X.-J. Wu, X. Zhu, and J. Kittler, "Exploring fusion strategies for accurate RGBT visual object tracking," *Inf. Fusion*, Article vol. 99, Nov 2023, Art no. 101881.

[14] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: A comprehensive review," *Inf. Fusion*, Review vol. 63, pp. 166-187, Nov 2020.

[15] 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.

[16] L. H. Huang, X. Zhao, and K. Q. Huang, "GOT-10k: A Large High-Diversity Benchmark for Generic Object Tracking in the Wild," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1562-1577, May 2021.

[17] G. Vivone, "Multispectral and hyperspectral image fusion in remote sensing: A survey," *Inf. Fusion*, Article vol. 89, pp. 405-417, Jan 2023.

[18] J. He *et al.*, "Spectral super-resolution meets deep learning: Achievements and challenges," *Inf. Fusion*, vol. 97, p. 101812, 2023/09/01/ 2023.

[19] 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.

[20] Z. Liu *et al.*, "An Anchor-Free Siamese Target Tracking Network for Hyperspectral Video," in *2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2021, pp. 1-5.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] Y. Zhang, X. Li, B. Wei, L. Li, and S. Yue, "A Fast Hyperspectral Tracking Method via Channel Selection," *Remote Sens.*, Article vol. 15, no. 6, Mar 2023, Art no. 1557.

[27] 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.

[28] 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.

[29] 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.

[30] C. Zhao, H. Liu, N. Su, and Y. Yan, "TFTN: A Transformer-Based Fusion Tracking Framework of Hyperspectral and RGB," *IEEE Trans. Geosci. Remote Sens.*, Article vol. 60, 2022 2022, Art no. 5542515.

[31] 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.*, pp. 1-1, 2023.

[32] 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.

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

[34] 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.

[35] B. Yan *et al.*, "Towards Grand Unification of Object Tracking," *in Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022.

[36] 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, 2018, vol. 11205, 2018, pp. 310-327.

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

[38] 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.

[39] 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.

[40] 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.

[41] Y. Su *et al.*, "Gaussian Information Entropy Based Band Reduction for Unsupervised Hyperspectral Video Tracking," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 2022, pp. 791-794.

[42] 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.

[43] Y. F. Li, C. J. Bian, and H. Z. Chen, "Object Tracking in Satellite Videos: Correlation Particle Filter Tracking Method With Motion Estimation by Kalman Filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 5630112.

[44] Y. Li, N. Wang, W. Li, X. Li, and M. Rao, "Object Tracking in Satellite Videos with Distractor-Occlusion Aware Correlation Particle Filters," *IEEE Trans. Geosci. Remote Sens.*, pp. 1-1, 2024.

[45] J. Kwon, H. S. Lee, F. C. Park, and K. M. Lee, "A Geometric Particle Filter for Template-Based Visual Tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 4, pp. 625-643, Apr 2014.

[46] W. L. Zheng, S. M. Bhandarkar, and Ieee, "A boosted adaptive particle filter for face detection and tracking," in *IEEE International Conference on Image Processing (ICIP 2006)*, Atlanta, GA, 2006, 2006, pp. 2821-+.

[47] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 5, pp. 564-577, May 2003.

[48] G. Y. Kulikov and M. V. Kulikova, "The Accurate Continuous-Discrete Extended Kalman Filter for Radar Tracking," *IEEE Trans. Signal Process.*, vol. 64, no. 4, pp. 948-958, Feb 2016.

[49] P. Zhang, J. Zhao, C. Bo, D. Wang, H. Lu, and X. Yang, "Jointly Modeling Motion and Appearance Cues for Robust RGB-T Tracking," *IEEE Trans. Image Process.*, Article vol. 30, pp. 3335-3347, 2021 2021.

[50] H. H. Nagel and W. Enkelmann, "An Investigation of Smoothness Constraints for The Estimation of Displacement Vector Fields from Image Sequences," *IEEE Trans Pattern Anal Mach Intell*, vol. 8, no. 5, pp. 565-93, May 1986.

[51] J. Shao, B. Du, C. Wu, and L. F. Zhang, "Tracking Objects From Satellite Videos: A Velocity Feature Based Correlation Filter," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7860-7871, Oct 2019.

[52] B. Du, S. H. Cai, and C. Wu, "Object Tracking in Satellite Videos Based on a Multiframe Optical Flow Tracker," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 8, pp. 3043-3055, Aug 2019.

[53] Y. Z. Chen, Y. Q. Tang, T. Han, Y. W. Zhang, B. Zou, and H. H. Feng, "RAMC: A Rotation Adaptive Tracker with Motion Constraint for Satellite Video Single-Object Tracking," *Remote Sens.*, vol. 14, no. 13, p. 3108, 2022.

[54] Z. Zhu, W. Wu, W. Zou, J. J. Yan, and Ieee, "End-to-end Flow Correlation Tracking with Spatial-temporal Attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, 2018, 2018, pp. 548-557.

[55] M. Danelljan, G. Bhat, S. Gladh, F. S. Khan, and M. Felsberg, "Deep motion and appearance cues for visual tracking," (in English), *Pattern Recognition Letters*, vol. 124, pp. 74-81, Jun 1 2019.

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

[57] 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.

[58] R. T. Collins, "Mean-shift blob tracking through scale space," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2003, vol. 2, pp. II-234.

[59] 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.

[60] 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.

[61] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale image Recognition," *in ICLR 2015*, 2014.

[62] Y. Wang *et al.*, "Spectral-Spatial-Aware Transformer Fusion Network for Hyperspectral Object Tracking," in *12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, Rome, ITALY, 2022.

[63] L. Zhao *et al.*, "Domain transfer and difference-aware band weighting for object tracking in hyperspectral videos," *Int. J. Remote Sens.*, Article vol. 44, no. 4, pp. 1115-1131, Feb 16 2023.

[64] R. Dian, S. Li, L. Fang, and Q. Wei, "Multispectral and hyperspectral image fusion with spatial-spectral sparse representation," *Inf. Fusion*, vol. 49, pp. 262-270, 2019/09/01/ 2019.

[65] J. Liu, Z. Wang, and M. Xu, "DeepMTT: A deep learning maneuvering target-tracking algorithm based on bidirectional LSTM network," *Inf. Fusion*, Article vol. 53, pp. 289-304, Jan 2020.

[66] S. Y. Xuan, S. Y. Li, M. F. Han, X. Wan, and G. S. Xia, "Object Tracking in Satellite Videos by Improved Correlation Filters With Motion Estimations," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1074-1086, Feb 2020.

[67] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *J. Basic Eng.*, vol. 82, no. 1, pp. 35-45, 1960.

[68] 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 7464-7473.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] G. Bhat, M. Danelljan, L. Van Gool, R. Timofte, and Ieee, "Learning Discriminative Model Prediction for Tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 6181-6190.
[86] 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.
[87] 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.
[88] 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.
[89] 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.
[90] 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.
[91] C. Mayer, M. Danelljan, D. Pani Paudel, and L. Van Gool, "Learning Target Candidate Association to Keep Track of What Not to Track," *in Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2021.
[92] D. Y. Guo *et al.*, "Graph Attention Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Electr Network, 2021, pp. 9538-9547.
[93] 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.
[94] 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.
[95] M. Paul, M. Danelljan, C. Mayer, and L. Van Gool, "Robust Visual Tracking by Segmentation," *in Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 571–588, 2022.
[96] B. Chen *et al.*, "Backbone is All Your Need: A Simplified Architecture for Visual Object Tracking," *in Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 375–392, 2022.
[97] F. Xie, C. Wang, G. Wang, Y. Cao, W. Yang, and W. Zeng, "Correlation-Aware Deep Tracking," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 8741-8750.
[98] S. Gao, C. Zhou, and J. Zhang, "Generalized Relation Modeling for Transformer Tracking," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18686-18695.
[99] X. Chen, H. Peng, D. Wang, H. Lu, and H. Hu, "SeqTrack: Sequence to Sequence Learning for Visual Object Tracking," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14572-14581.
[100] 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.
[101] 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.
[102] Y. B. Song *et al.*, "VITAL: Visual Tracking via Adversarial Learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8990-8999.

Author Statement

Yuzeng Chen: Conceptualization, Methodology, Validation, Formal analysis, Writing - Original Draft, Writing - Review & Editing.

Qiangqiang Yuan: Investigation, Methodology, Writing - Review & Editing, Funding acquisition.

Yuqi Tang: Resources, Data Curation, Writing - Review & Editing, Funding acquisition.

Yi Xiao: Writing - Review & Editing, Visualization, Methodology, Project administration.

Jiang He: Writing - Review & Editing, Visualization, Resources.

Zhenqi Liu: Writing - Review & Editing, Validation.

Declaration of interests

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

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: