SPIRIT: Spectral Awareness Interaction Network With Dynamic Template for Hyperspectral Object Tracking

Yuzeng Chen, Qiangqiang Yuan, *Member, IEEE*, Yuqi Tang, *Member, IEEE*, Yi Xiao, *Graduate Student Member, IEEE*, Jiang He, *Graduate Student Member, IEEE*, and Liangpei Zhang, *Fellow, IEEE*

***Abstract*— Hyperspectral (HS) video is able to capture abundant spectral, spatial, and temporal information about objects, which overcomes the limitations of common red-green-blue (RGB) video in complex scenarios such as similar appearances and background clutters (BCs). However, most trackers apply hand-crafted features extracted from manually selected bands instead of deep features for object representations due to limited HS data and the band gap problem. Each HS image consists of many bands, and it is challenging to fully interact with the band information while maintaining tracking speed. To this end, this article proposes a novel end-to-end spectral awareness interaction network with a dynamic template (SPIRIT) for HS video object tracking. First, a spectral awareness module (SAM) is proposed to learn band contributions with consideration of nonlinear and global interactions between HS bands. It can also cooperate with the feature extraction module pretrained with RGB data to attenuate the band gap and data-hungry. Second, an interaction module (IM) is proposed to achieve inter and intraband feature interactions to enhance tracking performance while improving efficiency. Furthermore, the proposed method contains a novel update module (UM) that evaluates the tracking confidence of the current state to adapt to object changes and attenuate tracking drifts. Extensive experiments demonstrate the superiority of our approach compared to state-of-the-arts (SOTAs) while meeting real-time demands.**

***Index Terms*— Dynamic template, hyperspectral (HS) object tracking, information interaction, spectral awareness.**

I. Introduction

VISUAL object tracking is an exciting research direction with a wide range of applications in autonomous driving, human-computer interaction, and augmented reality [1]. Given the initial state of an object, the objective is to locate its position and range in subsequent frames [2]. Most existing works [3], [4], [5], [6] are specialized for red-green-blue (RGB) video object tracking, which encounter difficulties in complex scenarios such as background clutters (BCs) and similar appearance [7], [8]. This is because the RGB image only contains three bands, which limits the tracker's ability in real-world scenes. With advancements in sensor technology, it is possible to obtain hyperspectral (HS) video that can record spectral, spatial, and temporal information [9], [10]. Especially, the rich spectral reflectance enables the tracker's potential material identification capability and provides more discriminative cues for object tracking in challenging cases [11]. However, achieving effective and efficient HS object tracking performance faces several key challenges, as follows:

1) *Data Hungry:* Deep learning techniques typically require a large number of training samples that are insufficient due to the sparsity of available HS samples [8]. This limitation makes it challenging to train accurate and generalized HS tracking models.
2) *Band Gap:* The number of bands in HS and RGB images does not coincide, which prevents directly unlocking the potential of the RGB model [7].
3) *Huge Volume:* HS video processing encounters high computational costs due to the existence of many narrow bands with high spectral correlation in each HS image.
4) *Arbitrary Change:* The appearance of the object often changes arbitrarily during tracking [12], and the lack of effective methods for adapting to these changes can leave trackers far from state-of-the-art (SOTA).

HS trackers typically rely on robust features to distinguish objects from the background. Some HS trackers, such as [13], [14], and [15] use hand-crafted features, while the discriminative capability of these features is limited compared to deep features learned from HS datasets. Moreover, hand-crafted features may not generalize well to arbitrary types of objects. Recent advances in RGB data processing have demonstrated the effectiveness of deep networks [4], [16], [17], [18], [19], [20]. Considering the data-hungry, it is challenging to train an HS tracking network due to the sparse training samples. For this reason, large-scale datasets in the RGB tracking field such as GOT-10K [21], TrackingNet [22], and LaSOT [23] can be reused to train HS networks [8]. Nevertheless, bridging the gap between pretrained RGB models and HS networks is demanding due to the different number of bands in RGB

Manuscript received 19 October 2023; revised 23 November 2023; accepted 16 December 2023. Date of publication 28 December 2023; date of current version 9 January 2024. This work was supported by the National Natural Science Foundation of China under Grant 42230108 and Grant 61971319. *(Corresponding author: Qiangqiang Yuan.)*

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

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

Liangpei Zhang is with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan 430079, China (e-mail: zlp62@whu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2023.3347950

and HS images [7]. To attenuate this issue, some methods have been proposed to convert the HS image to a three-band image and perform subsequent tasks. This can be achieved by selecting three bands directly or using dimension-reduction techniques to obtain a false-color image, such as [24], [25], and [26].

However, using only three bands can lead to a loss of spectral information. Therefore, it is considered to either repeat each band three times or group adjacent three bands to generate multiple false-color images for feeding into a pretrained RGB network, which suffers from data redundancy and violates the fact that the information in red, green, and blue channels is usually inconsistent in read world. In addition, not all bands contribute equally to the downstream tracking task, and treating them equally can lead to suboptimal performance [7]. Recent works, such as SEE-Net [7], BAE-Net [27], SST-Net [28], and BRRF-Net [29], have been developed and demonstrated excellent performance. Some of these trackers, such as SST-Net [28] and BRRF-Net [29], directly apply average ensemble learning to fuse weak tracking results or response maps, which may cause track drifts due to equalizing different false-color images. While the SEE-Net [7] has shown competitive performance in HS tracking by dynamically aggregating weak results based on the importance of false-color images. However, the efficiency of the average ensemble learning fashion could be limited due to the need to predict multiple weak tracking results in each inference session. Moreover, to adapt to object changes, some effective strategies, such as [7], [12], [28], [30], and [31], have been proposed. For example, SiamF [30] incorporates an online material classifier to consider object changes, increasing tracking robustness. The template in Siamese trackers is typically initialized in the first frame and is kept fixed or slightly updated for the rest of the HS video. This can lead to tracking drift over time, especially when there are frequent appearance changes. Therefore, updating the template is essential for maintaining accurate HS tracking results [12].

Based on the above analysis, it can be found that most HS trackers employ hand-crafted features captured from manually selected bands for representing HS objects, which may limit the robustness and automaticity of algorithms. In particular, each image in HS video consists of many bands, and it is always challenging to fully interact with the band information while maintaining a real-time running speed. Additionally, the lack of strategies to handle HS object changes may also elevate the risk of tracking drifts. In this article, we propose an efficient end-to-end network called the spectral awareness interaction network with a dynamic template (SPIRIT) for HS object tracking. SPIRIT consists of five modules: spectral awareness module (SAM), feature extraction module, interaction module (IM), prediction module, and update module (UM). To handle the band gap between HS and RGB images, we propose a SAM that learns the relationship between bands to evaluate their contributions to downstream tasks. This module accounts for nonlinear and global interactions between spectral bands from the perspective of spectral reconstruction. Guided by the band contributions, the HS image is adaptively divided into multiple false-color images with different contributions and low correlation. These images are then fed to the feature extraction module to obtain a deep visual representation. To attenuate the data-hungry issue and fully exploit pretrained RGB models, the feature extraction module in a transferred tracking network pretrained with RGB data is utilized to construct our HS tracker. The SAM and feature extraction module can adaptively extract deep features of HS images while retaining the discriminative capability learned from massive RGB data. To acquire inter and intraband interaction (IBI) information and maintain high efficiency, a novel IM is proposed. More concretely, it first performs the interband feature interaction, and an IBI network is then proposed to learn IBI information. The learned inter and intraband features are fused and sent to the prediction module for localization. Moreover, we integrate a UM into the SPIRIT, enabling us to acquire spectral, spatial, and temporal information to adapt to object changes. The main contributions are summarized as follows:

1) A SAM is proposed to learn the band contributions with consideration of nonlinear and global interactions between HS bands. The SAM cooperates with the feature extraction module pretrained with RGB data to attenuate the band gap and data-hungry.
2) An IM is proposed for achieving inter and intraband feature interactions to improve tracking accuracy while ensuring efficiency. Furthermore, our proposed approach includes a novel UM that evaluates the HS tracking confidence of the current state to adapt to object changes to attenuate tracking drifts.

Comprehensive comparisons and intensive analyses are performed to demonstrate the superiority of the proposed SPIRIT method in terms of tracking effectiveness and efficiency, as shown in Fig. 1. The rest of this article is organized as follows. Section II reviews the related work including RGB trackers and HS trackers. The proposed approach is introduced in Section III. Experimental results and analysis are presented in Section IV. In Section V, we conclude the article and summarize the contributions.

II. Related Work

HS video object tracking shows great potential in compensating for the limitations of current RGB trackers in complex scenarios such as BC and similar appearances. This section provides an overview of related research works on both RGB and HS trackers.

A. Visual Object Tracking in RGB Videos

In general, visual trackers can be classified into two main categories: generative paradigm and discriminative paradigm [32]. The generative paradigm involves constructing a model manually to represent the object and then finding a region that is similar to the description of the generative model by classifying the signal and minimizing the objective loss. The accuracy and speed of trackers are directly affected by object representation models, such as kernel trick [33], sparse representation [34], and Gaussian mixed model [35].

In the discriminative paradigm, the correlation filter and the Siamese network are two popular examples [2]. Owing

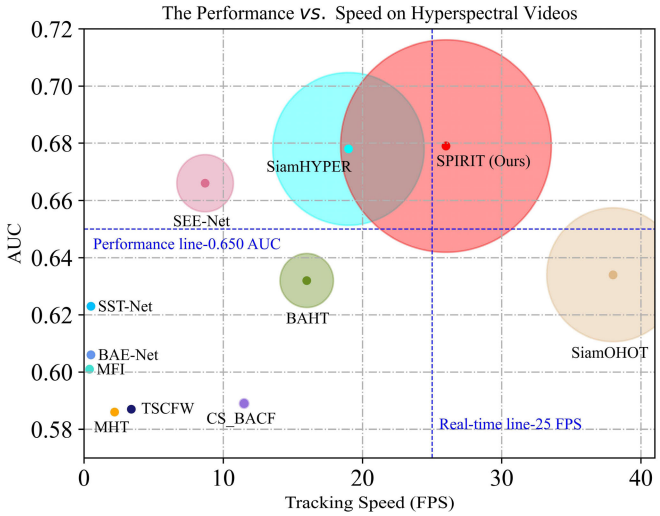


Fig. 1. Comparison with SOTA HS trackers. We visualize the AUC versus tracking speed in terms of FPS. The size of the circle represents the weighted sum of the tracking speed (x-axis) and AUC (y-axis). A larger circle indicates better performance.

to their simplicity and effectiveness, they have become a fundamental paradigm in recent decades. Some efforts, such as [36], [37], [38], [39], and [40] have been made in correlation filters with hand-crafted features such as intensity, color name [41], and histogram of oriented gradients [42]. By minimizing a least-squares error, the correlation filter learns a classifier to determine the object state and updates the model to adapt to object changes. However, hand-crafted features have limitations in object representations. In contrast, shallow convolutional features contain low-level information with high spatial resolution, which is suitable for accurate localization, while deeper features encode high-level information with low resolution (LR), promising to improve tracking robustness. Driven by deep learning, many elaborate trackers such as [5], [17], [43], and [44] inherit the Siamese network. These models usually consist of two branches: the template branch and the search branch. The template branch takes a patch of the first frame as input, while the search branch receives patches from subsequent frames. Both branches share a network trained from massive samples to ensure that the same transformation is imposed on these two branches [2]. Although Siamese trackers exhibit superior performance, they still face challenges such as BC and similar appearance due to limited spectral information in RGB images [7], [8].

B. Visual Object Tracking in HS Videos

With the advancement of imaging technology, HS cameras can capture rich spectral, spatial, and temporal information simultaneously [45], [46]. In particular, the spectral information can reflect the material properties of objects, making HS videos promising for tracking objects in complex scenarios [7], [11]. Several excellent methods have been developed for HS video object tracking, including both generative and discriminative paradigms. Early efforts are based on the generative paradigm, which determines the object by designing a representation model. For example, Banerjee et al. [47] and Hien Van et al. [48] explore the spectral angle mapper and mean shift [49] to achieve HS object tracking.

For the discriminative paradigm, HS trackers jointly use foreground and background regions to determine the object state, thus improving tracking performance. Specifically, HS trackers, such as [11], [13], [14], and [15], inherit correlation filters and aim to exploit full-band spectral features. For example, material based hyperspectral tracker (MHT) [11] studies the tracking task from the perspective of material features by using abundance features and local spectral-spatial histograms of multidimensional gradients. While tensor-based sparse correlation filter with spatially and spectrally weighted and regularized (TSCFW) [15] studies tensor processing to reduce spectral differences in homogeneous backgrounds. Meanwhile, the sparse regularization term and context-aware information are integrated into the correlation filter to suppress false responses. The results have verified their effectiveness. Robust features are the basis for reliable tracking, while discriminative models determine the tracking performance [2]. Compared to traditional correlation filters, Siamese-based HS trackers, such as [7], [8], [25], [31], [50], [51], and [52], are more discriminative because they can leverage well-trained RGB models to learn a common HS object representation. For example, SEE-Net [7] first divides the HS image into multiple false-color images that are then transferred to a SOTA RGB tracker, and the final state is obtained by embedding multiple weak tracking results. Similarly, BAE-Net [27] and SST-Net [28] integrate multiple weak tracking results, which may result in an expensive computational burden. Additionally, object changes in appearance pose a significant challenge, and effective strategies have been proposed to solve this issue such as [7], [12], [28], [30], and [31]. However, the Siamese-based HS trackers usually use the first frame to initialize tracking template and remain fixed or slightly updated in the subsequent frames of HS videos, making it difficult to exploit the temporal information of objects.

III. Proposed Approach

In this section, we detail the proposed HS tracking method, which includes the overall architecture, SAM, IM, prediction module, and UM, as well as the training and inference processes.

A. Overall Architecture

As illustrated in Fig. 2, the proposed SPIRIT HS tracker is composed of five modules: the SAM, feature extraction module, IM, prediction module, and UM. The input is a triple consisting of the initial template, search region, and dynamic template. As mentioned earlier, the band gap prevents bridging pretrained RGB models to HS trackers directly. To attenuate this issue, we propose a SAM that takes into account the nonlinear and global interactions between spectral bands from the perspective of spectral reconstruction. Meanwhile, it learns the relationship between bands to evaluate their contributions, allowing the HS image to be adaptively grouped into multiple false-color images. The feature extraction module in a transferred RGB network is employed to extract deep features of

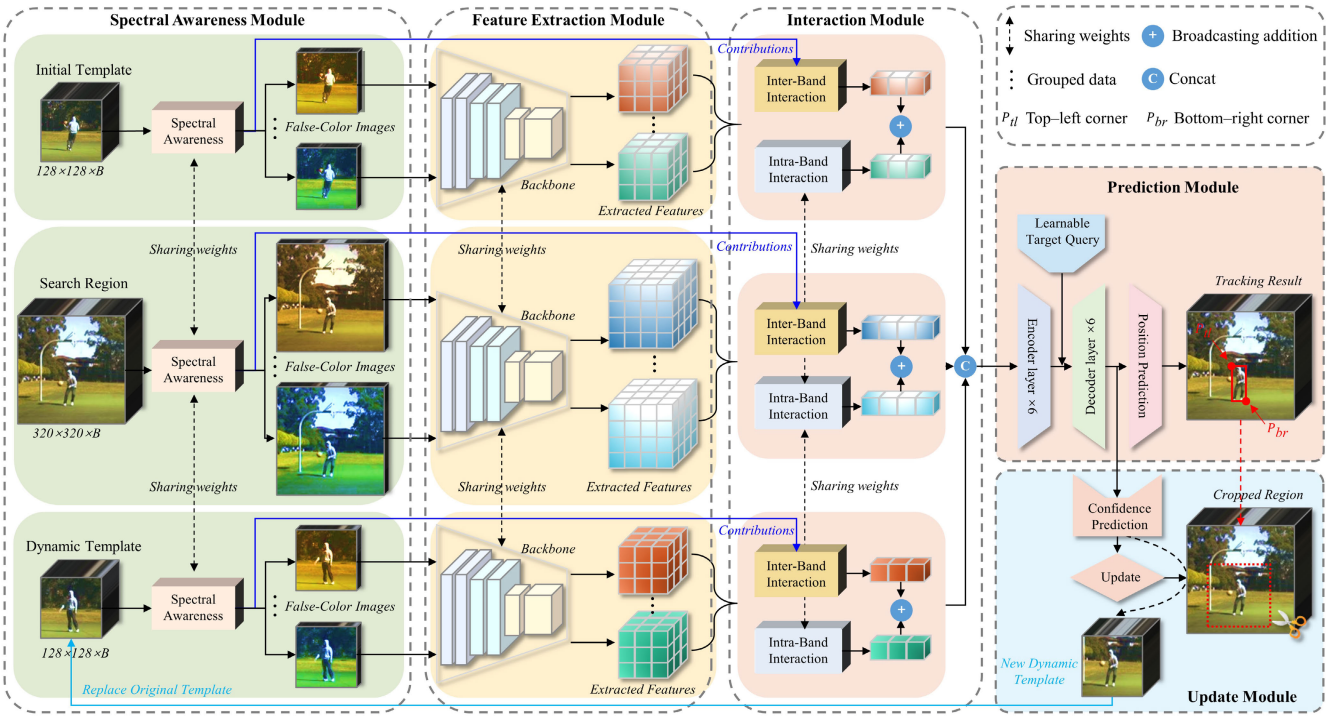


Fig. 2. Overall architecture of the proposed SPIRIT HS tracker. The SPIRIT is a Siamese-based HS tracker composed of five modules: SAM, feature extraction module, IM, prediction module, and UM.

HS bands while obtaining a discriminative object representation. To further achieve inter and intraband feature interactions, the deep features generated from false-color images are first integrated to realize the interband interaction based on learned band contributions. Then, we propose an IBI network to learn the IBI features. The interband and IBI features are flattened, summed, and concatenated to feed into the prediction module for localization. Furthermore, the proposed method includes a novel UM that evaluates the tracking confidence of the current state to adapt to object changes in appearance. In the UM, a dynamically updated template is embedded to simultaneously leverage the spectral, spatial, and temporal information for handling tracking drifts in HS object tracking.

B. Spectral Awareness Module

To attenuate the band gap, we propose a SAM that converts the HS image into multiple false-color images by exploiting the spectral and spatial information. The structure of the SAM is shown in Fig. 3, which consists of two parts: band excitation and band reconstruction. The band excitation is designed to calculate spectral contributions while the band reconstruction part reconstructs the HS image to optimize the contribution by minimizing the reconstruction error.

For an HS video, each frame I is denoted as $I \in \mathbb{R}^{M\times N\times B}$ with $M \times N$ pixels and B bands. I can be seen as a band set $I = [b_1, b_2, \ldots, b_B]$, where b_i denotes the ith band. First, I is input to the SAM to seek the interdependence among bands by a function h. Then, we can obtain the spectral contributions of all bands by

$$Z = \sigma h(I; \alpha) \tag{1}$$

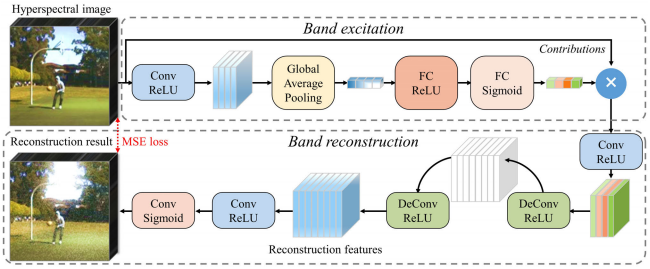


Fig. 3. Structure of the SAM. It consists of the band excitation and band reconstruction parts. Conv denotes the standard convolution. FC is a fully connected layer. DeConv is a fractionally strided convolution layer. ReLU and Sigmoid are the rectified linear unit activation and Sigmoid activation.

where $Z \in \mathbb{R}^{1\times 1\times B}$ is the band contribution with $Z = [z_1, z_2, \ldots, z_B]$, σ is the Sigmoid activation, and α denotes the learnable parameter of h. Furthermore, we apply the multiplication operation to establish the excitation result between the raw HS image I and band contribution Z, as follows:

$$R = I \otimes Z \tag{2}$$

where $\otimes$ represents the element-wise multiplication and R refers to the result of band excitation.

To understand the interdependence among bands, it is required to reconstruct the original HS image from the excitation result R. The reconstruction function g with learnable parameters β takes R as input to obtain predicted reconstruction result $\hat{I}$, as follows:

$$\hat{I} = g(R; \beta). \tag{3}$$

The mean squared error (MSE) is used to measure the reconstruction performance. The loss function is formulated

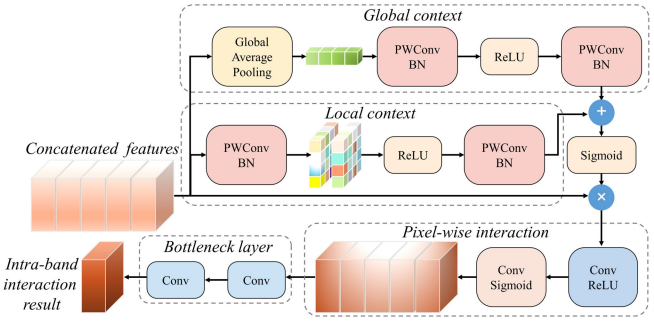


Fig. 4. Structure of the IBI network in IM. The network takes the concatenated features as input and achieves channel feature interaction with global and local contexts. Moreover, a pixel-wise features interaction is implemented followed by a bottleneck layer to obtain the final IBI result. PWConv denotes the point-wise convolution.

as follows:

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{B} \left\| b_i - \hat{b}_i \right\|^2 + \lambda \|\alpha\beta\|^2 \tag{4}$$

where $\mathcal{L}_{\text{rec}}$ represents the reconstruction loss, B is the number of bands, and $\hat{b}_i$ is the ith band of the predicted reconstruction result $\hat{I}$. $\lambda\|\alpha\beta\|^2$ is used to limit the complexity of the model and maintain the sparsity of band contributions [53]. α and β can be solved by minimizing the loss $\mathcal{L}_{\text{rec}}$ in a gradient descent fashion. So far, we are able to obtain the band contribution that interprets the interdependence among spectral bands.

Benefiting from the end-to-end training fashion, the SAM is able to cooperate with downstream modules to develop a supervised scheme for learning the band contribution. Therefore, the SAM not only produces the band contribution but also serves downstream modules. Guided by band contributions Z, we sort all bands to form $n = \text{int}(B/3)$ false-color images $D = [d_1, d_2, \ldots, d_n]$. The contribution w of each false-color image is computed by summing the contributions of each band of each false-color image and then dividing by the number of bands. All contributions of false-color images can be expressed by $W = [w_1, w_2, \ldots, w_n]$. Finally, the generated false-color images D are fed to the feature extraction module, and the contribution W will also be reused to achieve interband feature interaction in the IM.

C. Interaction Module

The band awareness module yields false-color images D and their contributions W. Then, these false-color images are fed to the feature extraction module, which is a modified ResNet [54] network with the last stage and fully connected layer removed. To exploit the full-band information and maintain high tracking efficiency, we propose an IM (see Fig. 4) that takes the features extracted from the initial template, dynamic template, and search region as input for inter and intraband feature interactions, shown in Fig. 2.

As discussed above, the contribution W records the impact of false-color images on downstream tracking tasks and can be utilized to adaptively aggregate extracted features for robust HS object tracking. Specifically, the interband feature interaction takes as input deep features generated by false-color images and aggregates them by embedding learning fashion, as follows:

$$Q = \frac{\sum_{i=1}^{n} w_i \varphi(d_i)}{\sum_{i=1}^{n} w_i} \tag{5}$$

where d_i denotes the ith false-color image, w_i denotes the contribution of d_i, φ expresses the feature extraction function, n is the number of false-color images, and Q denotes the interband interaction result. It is worth noting that w_i varies with the input HS image, so that the interband interaction can adaptively integrate features for tracking.

However, the interband interaction ignores the information across all bands. To attenuate this problem, an IBI called IBI network is proposed, as shown in Fig. 4. It takes a concatenated result of the $\varphi(d_1), \varphi(d_2), \ldots, \varphi(d_n)$ as input and obtains IBI features. Inspirited by the attention mechanism, the IBI network first embeds channel features at both global and local scales followed by adaptively mining the pixel features of different regions. Therefore, the IBI network is able to provide additional flexibility for feature interactions across all spectral bands of the HS image.

The channel attention mechanism usually squeezes each channel into one value. This coarse manner tends to emphasize the global feature texture. However, HS images usually have a large percentage of backgrounds, and the global context focuses more on backgrounds and eliminates most of the band signals of interest objects. Suboptimal results would be achieved if only the global channel attention mechanism is available in IBI. To solve this issue, we synergize the local with global contexts for information interaction.

Given concatenated feature $X \in \mathbb{R}^{H\times W\times C}$ with C channels and $H \times W$ pixels, the global context is obtained by

$$G(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{[i,j,:]} \tag{6}$$

$$\text{global}(X) = \mathcal{B}(PW_2\delta(\mathcal{B}(PW_1(G(X))))) \tag{7}$$

where $G(\cdot)$ denotes the global average pool (GAP), $\text{global}(\cdot) \in \mathbb{R}^{1\times 1\times C}$ denotes the global feature, $\mathcal{B}$ is the batch normalization (BN), δ is the rectified linear unit (ReLU) activation function, and PW_1 and PW_2 are two point-wise convolutions.

For the local context feature, we perform the IBI of each spatial location

$$\text{local}(X) = \mathcal{B}(\text{PW}_4(\delta(\mathcal{B}(\text{PW}_3(X))))) \tag{8}$$

where $\text{local}(\cdot) \in \mathbb{R}^{H\times W\times C}$ denotes the local context. It is noted that the width and height of $\text{local}(\cdot)$ are equal to the sizes of input features, preserving, and emphasizing detailed features across overall bands.

So far, we have obtained the global context $\text{global}(X)$ and local context $\text{local}(X)$ along channel dimensions, and they are refined by

$$X' = X \otimes \sigma(\text{global}(X) \oplus \text{local}(X)) \tag{9}$$

where $\oplus$ denotes the broadcasting addition. $X' \in \mathbb{R}^{H\times W\times C}$ stands for the refined feature that takes the global and local contexts into account for intraband feature interactions. However, the refined feature X' pays more attention to

channel-wise interactions. Hence, pixel attention is imposed so that it focuses more on specific object regions, i.e., the spatial dimension. Specifically, the feature X' is fed to two standard convolutions (i.e., Conv_1 and Conv_2) with ReLU activation and Sigmoid activation, as follows:

$$Y = X' \otimes \sigma\left(\text{Conv}_2\left(\delta\left(\text{Conv}_1\left(X'\right)\right)\right)\right) \tag{10}$$

where Y denotes the result of the pixel attention interaction.

Finally, the output F of the IBI network is obtained by

$$F = \text{Conv}_4(\text{Conv}_3(Y)) \tag{11}$$

where Conv_1 and Conv_2 compose the bottleneck layer for dimension reduction and obtaining a more compact output. The IBI network encodes the channel-wise feature and pixel-wise feature for intraband feature interaction, enhancing object representations of tracked HS objects.

The result of the IM is obtained by flattening and summing interband and IBI features. The interacted features of the initial template, dynamic template, and search region are denoted as F_{init}, F_{dy}, and F_{sear}, respectively. They are concatenated to generate a feature sequence followed by delivering it to the prediction module for localization.

D. Prediction and Update Modules

1) Prediction Module: In Fig. 2, the body of the prediction module is an encoder-decoder transformer structure (see Fig. 5) followed by a position prediction part. The encoder is composed of six encoder layers, and each encoder layer consists of a multihead self-attention coupled with a feed-forward neural network (FFN), as shown in Fig. 5(a). The encoder can learn the dependencies across spectral, spatial, and temporal features in the HS sequence, therefore improving the model's discriminative capability. The decoder receives a learnable target query and a sequence of features from the encoder to predict the desired bounding box. The decoder consists of six decoder layers. In contrast to the encoder, each decoder layer is composed of a multihead self-attention, a multihead cross-attention, and an FFN, as shown in Fig. 5(b). In multihead cross-attention, a target query is trained to focus on all positions on the template and search region features [55], yielding robust object representations for localization. For the bounding box prediction, it takes the search region features from the encoder and the output of the decoder as input. The salient features in search region are then enhanced by an attention mechanism. The new features are reshaped and fed to the fully convolutional network to predict probability maps of the bottom-right and top-left corners of the object. Finally, the object's coordinates are computed by

$$\begin{aligned}\left(\hat{x}_{\text{br}}, \hat{y}_{\text{br}}\right) &= \left(\sum_{x=0}^{W'}\sum_{y=0}^{H'} x\,P_{\text{br}}(x,y), \sum_{x=0}^{W'}\sum_{y=0}^{H'} y\,P_{\text{br}}(x,y)\right)\\ \left(\hat{x}_{\text{tl}}, \hat{y}_{\text{tl}}\right) &= \left(\sum_{x=0}^{W'}\sum_{y=0}^{H'} x\,P_{\text{tl}}(x,y), \sum_{x=0}^{W'}\sum_{y=0}^{H'} y\,P_{\text{tl}}(x,y)\right)\end{aligned} \tag{12}$$

where $(\hat{x}_{\text{br}}, \hat{y}_{\text{br}})$ and $(\hat{x}_{\text{tl}}, \hat{y}_{\text{tl}})$ indicate the predicted bottom-right and top-left corners and $P_{\text{br}}(x, y)$ and $P_{\text{tl}}(x, y)$ denote corresponding probability maps with a size of $H' \times W'$ pixels.

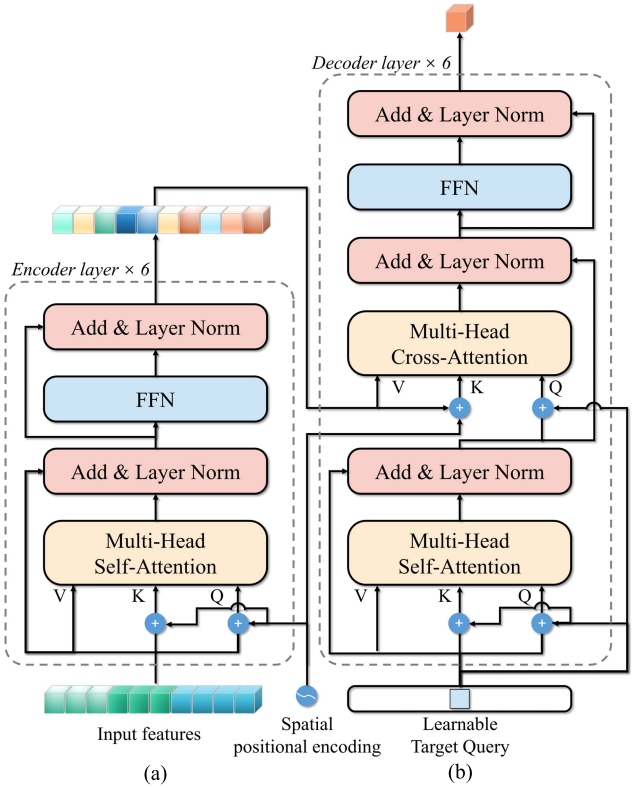


Fig. 5. Structure of the Transformer encoder-decoder in the prediction module. (a) Encoder takes input features and a sinusoidal positional embedding as input. (b) Decoder receives the output of the encoder and a learnable target query to obtain the final result.

2) Update Module: As discussed above, for Siamese-based HS trackers, the template is usually initialized in the first frame and kept fixed or slightly updated in subsequent frames. Nevertheless, object appearance often varies from frame to frame, and failure to update the template in time could lead to tracking drifts, especially for objects in long time-series sequences. To handle this issue, the proposed SPIRIT tracker cooperates with an efficient yet effective UM in which a dynamically updated template sampled from intermediate frames is treated as an auxiliary input, as shown in Fig. 2. The UM takes the output of the decoder as input and aims to evaluate the tracking confidence of current state for updating the template dynamically. The confidence evaluation function is implemented by a three-layer perceptron network followed by a Sigmoid activation. If the tracking confidence is higher than a common threshold $\tau = 0.5$ and the interval is reached, the tracking state will be considered reliable and the original dynamic template will be updated by a new dynamic template (see Fig. 2). In brief, the UM can attach temporal information to HS object tracking task, thus adapting to object changes in appearances over time.

E. Training and Inference

1) Training: Considering limited HS samples, initializing the parameters of SPIRIT in a random fashion would achieve unsatisfactory effects. To this end, the feature extraction

module, prediction module, and UM are initialized with network parameters pretrained on RGB data. While parameters of the SAM and IM are initialized in a random manner. In training, we train the parameters of SAM, IM, and UM, and the remaining modules are frozen to better unlock the potential of pretrained RGB models. We sample a triple including the initial template, dynamic template, and search region to train the proposed SPIRIT tracker. The sizes of the triple are $128 \times 128 \times B$, $128 \times 128 \times B$, and $320 \times 320 \times B$, respectively. The proposed SPIRIT tracker is trained with multitask losses in an end-to-end fashion, as follows:

$$\mathcal{L}_{\text{total}} = \gamma_{\text{rec}}\mathcal{L}_{\text{rec}} + \gamma_{l1}\mathcal{L}_{l1} + \gamma_{\text{diou}}\mathcal{L}_{\text{diou}} + \gamma_{\text{cl}}\mathcal{L}_{\text{cl}} \tag{13}$$

where $\mathcal{L}_{\text{rec}}$ represents the reconstruction loss (1) of SAM, $\mathcal{L}_{l1}$ denotes the $L1$ loss, and $\mathcal{L}_{\text{diou}}$ denotes distance intersection over union loss [56] between the ground-truth and predicted box in the prediction module, and $\mathcal{L}_{\text{cl}}$ denotes the cross-entropy loss in UM. γ_{rec}, γ_{l1}, γ_{diou}, and γ_{cl} are the weight factors.

For $\mathcal{L}_{\text{rec}}$, it is the sum of the triple reconstruction losses, i.e., the initial template loss $\mathcal{L}_{\text{rec}}^{\text{init}}$, search region loss $\mathcal{L}_{\text{rec}}^{\text{sear}}$ and dynamic template loss $\mathcal{L}_{\text{rec}}^{\text{dymc}}$, as follows:

$$\mathcal{L}_{\text{rec}} = \frac{1}{2}\left(\mathcal{L}_{\text{rec}}^{\text{sear}} + \frac{1}{2}\left(\mathcal{L}_{\text{rec}}^{\text{init}} + \mathcal{L}_{\text{rec}}^{\text{dymc}}\right)\right). \tag{14}$$

For $\mathcal{L}_{\text{diou}}$, it is defined as

$$\mathcal{L}_{\text{diou}} = 1 - \text{IoU}\left(B_i, \hat{B}_i\right) + \frac{\rho^2\left(b_i, \hat{b}_i\right)}{c^2} \tag{15}$$

where B_i and $\hat{B}_i$ denote the ground truth and predicted box, b_i and $\hat{b}_i$ denote the centers of B_i and $\hat{B}_i$. $\rho(\cdot)$ represents the Euclidean distance, and c represents the diagonal length of the smallest enclosing box that covers B_i and $\hat{B}_i$.

For the $\mathcal{L}_{\text{cl}}$, it is defined as

$$\mathcal{L}_{\text{cl}} = y_i \log(l_i) + \left(1 - y_i\right)\log(1 - l_i) \tag{16}$$

where l_i denotes the predicted tracking confidence of the current state and y_i denotes the ground-truth label.

2) Inference: In the inference process, we first crop the initial and dynamic templates from the first HS frame and divide them into two sets of false-color images based on the SAM, as shown in Fig. 2. These false-color images are then fed to the feature extraction module to extract deep features. The IM takes corresponding features to realize inter and intraband feature interactions. Similarly, for subsequent frames, a search region is cropped from the HS frame and sequentially fed to the SAM, feature extraction module, and IM. Then, the interacted features of the initial template, dynamic template, and search region are concatenated and fed to the prediction module to yield the final bounding box. In addition, the output of the decoder is fed to the UM to compute the tracking confidence for updating the dynamic template accordingly. The new dynamic template will also be cropped from the current HS image to replace the original template, adapting to object changes in appearance over time.

IV. Experiments

A. Experimental Setups

1) Data Description: The proposed method is trained and tested on the HS Object Tracking Competition dataset [11]. The dataset consists of 40 sets of training videos and 35 sets of test videos, totaling 75 sets of videos, captured at 25 frames/s. Each set of videos contains three types of data, i.e., HS video (16 bands), false-color video (three bands), and RGB video (three bands). The false-color videos are generated by corresponding HS videos. The RGB video is captured from a viewpoint close to the HS video. The dataset is annotated with 11 finely attributes including scale variation (SV), fast motion (FM), BC, in-plane rotation (IPR), out-of-plane rotation (OPR), occlusion (OCC), deformation (DEF), motion blur (MB), illumination variation (IV), LR, and out-of-view (OV). Each object is labeled with a horizontal bounding box indicated by the center position, width, and height. HS and false-color videos share labels, while RGB video's labels are generated independently.

2) Implementation Details: The proposed approach is implemented in Python with PyTorch and trained on a machine equipped with the Intel Core i7-12700F Central Processing Unit (CPU) and NVIDIA GeForce RTX 4060 Graphics Processing Unit (GPU). The initial network [55] is pretrained on RGB datasets consisting of train-splits of the GOT-10K [21], TrackingNet [22], LaSOT [23], and COCO [57]. Subsequently, the HS dataset is further utilized for training by using the Adam with decoupled weight decay (AdamW) method [58] with a learning rate of $5e^{-5}$ and weight decay of $1e^{-4}$. The values for γ_{rec}, γ_{l1}, γ_{diou}, and γ_{cl} are implicitly set to 0.012, 5.0, 2.0, and 0.5, respectively. The number of heads in the multihead attention is set to 8, and the dropout rate is set to 0.1. The training sizes of the search and template patches are set to $320 \times 320 \times 16$ pixels and $128 \times 128 \times 16$ pixels, respectively. The network is then trained for a total of 20 epochs with a batch size of 8.

3) Evaluation Metrics: Success plots and precision plots are applied to benchmark the tracking effects in a one-pass evaluation fashion [59]. In the success plot, overlap is commonly used for evaluation. Given the ground truth R_G and predicted result R_T, the overlap is calculated by

$$\text{overlap} = \frac{|R_G \cap R_T|}{|R_G \cup R_T|} \tag{17}$$

where $\cap$ and $\cup$ denote the intersection and union operators respectively, and $|\cdot|$ computes the number of pixels in the region. The success plot indicates that the success rate exceeds the threshold range $T_s \in [0, 1]$.

The precision plot displays the percentage of frames where the center location error (CLE) is smaller than the predefined threshold $T_k \in [1, 50]$. CLE, a common measure of distance, is calculated by

$$\text{CLE} = \sqrt{(\text{CT}_x - \text{GT}_x)^2 + \left(\text{CT}_y - \text{GT}_y\right)^2} \tag{18}$$

where $(\text{CT}_x, \text{CT}_y)$ and $(\text{GT}_x, \text{GT}_y)$ are the centers of the R_T and R_G, respectively. The precision plot evaluates the performance in localization while the success plot accounts for

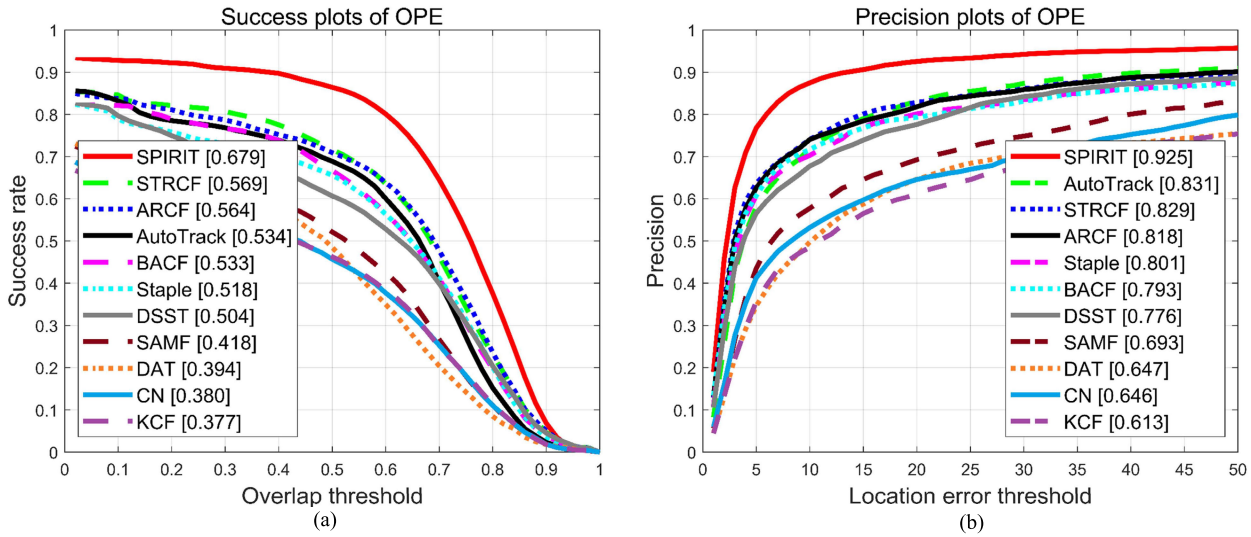


Fig. 6. Comparison with RGB trackers using hand-crafted features on RGB videos. (a) Success plot. (b) Precision plot.

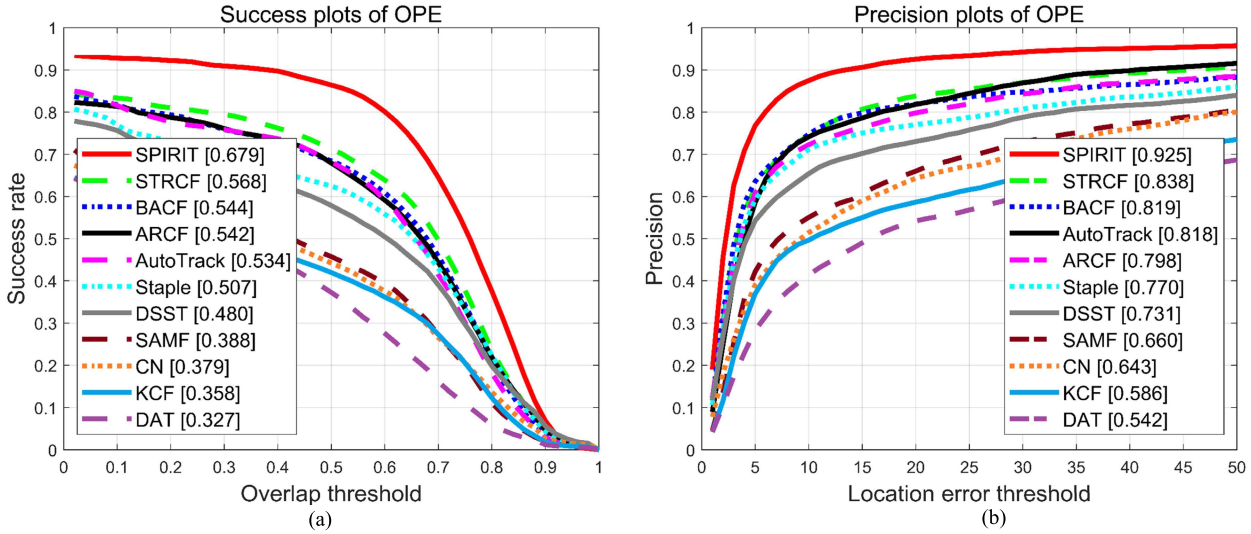


Fig. 7. Comparison with RGB trackers using hand-crafted features on false-color videos. (a) Success plot. (b) Precision plot.

both localization and scale estimation performance of trackers. The area under the curve (AUC) of the success plot and the distance precision (DP) at 20 pixels of the precision plot are used for evaluation. In this article, we primarily use AUC to rank all trackers, and the frames per second (FPS) is utilized to measure the running efficiency.

B. Quantitative Evaluation With RGB Trackers

1) Hand-Crafted Feature-Based Trackers: We compare the proposed SPIRIT tracker with ten SOTA trackers using hand-crafted features, namely AutoTrack [3], kernelized correlation filter (KCF) [37], DSST [38], Staple [39], CN [41], ARCF [60], BACF [61], spatial-temporal regularized correlation filter (STRCF) [62], SAMF [63], and DAT [64]. These trackers are tested on RGB and false-color videos, while SPIRIT is tested on HS videos. Fig. 6 displays the success and precision plots tested on RGB videos, and the legends indicate the AUC and DP values, respectively.

Notably, the proposed SPIRIT achieves the optimal results, with an AUC of 0.679 and DP of 0.925, while KCF performs the worst. Among the compared trackers, STRCF ranks first by balancing aggressive and passive learning to adapt to significant object changes in appearance. Compared with STRCF, the SPIRIT yields a gain of 11.0% in AUC and 9.6% in DP. Furthermore, compared to ARCF and AutoTrack, SPIRIT improves the AUC by 11.5% and 14.5%, respectively, highlighting the benefits of utilizing spectral, spatial, and temporal features. These results suggest that the rich spectral and temporal information present in HS videos can enhance tracking effects.

Intuitively, HS videos can be converted to false-color videos followed by using RGB trackers for object tracking. Fig. 7 illustrates the results tested on generated false-color videos. Remarkably, STRCF still achieves a respectable performance, with an AUC of 0.568 and a DP of 0.838. The proposed SPIRIT tracker demonstrates superior performance, achieving

TABLE I
COMPARISON WITH DEEP FEATURE-BASED TRACKERS IN TERMS OF AUC

	ECO [65]	SiamCAR [5]	SiamBAN [66]	Stark [55]	LightTrack [67]	DaSiamRPN [68]	DiMP [69]
RGB	0.577	0.636	0.610	0.637	0.593	0.622	0.641
HS/F	0.556	0.586	0.587	0.579	0.530	0.575	0.556
	ATOM [19]	RTS [70]	SiamGAT [71]	SiamRPN++ [17]	SimTrack [4]	SPIRIT (Ours)	
RGB	0.614	0.612	0.649	0.653	0.664	n/a	
HS/F	0.556	0.568	0.576	0.591	0.602	0.679	

RGB, HS and F denote the red-green-blue, hyperspectral, and false-color videos, respectively. The first, second, and third places are highlighted in red, blue, and magenta, respectively. n/a denotes not applicable.

11.1% and 8.7% improvements in AUC and DP, respectively. Experimental results underscore the suboptimal nature of converting HS videos to false-color videos for object tracking, as it inevitably leads to a loss of material spectral information that is crucial for achieving robust tracking performance.

2) Deep Feature-Based Trackers: We conduct a comprehensive comparison with 12 SOTA trackers that utilize deep features, including SimTrack [4], SiamCAR [5], SiamRPN++ [17], ATOM [19], Stark [55], ECO [65], SiamBAN [66], LightTrack [67], DaSiamRPN [68], DiMP [69], RTS [70], and SiamGAT [71]. These trackers encompass diverse types of backbones and tracking paradigms. Table I presents the experimental results tested on RGB and false-color videos.

It is worth remarking that SimTrack, a unified transformer tracking framework, achieves outstanding results, yielding an AUC of 0.664 and 0.602 on RGB and false-color videos, respectively. Compared with SimTrack, the proposed SPIRIT gains by 1.5% and 7.7%, respectively. Additionally, Stark, an anchor-free transformer tracker that can capture long-range dependencies in both spatial and temporal dimensions, is surpassed by the SPIRIT by 4.2% and 10.0% on RGB and false-color videos, respectively. Overall, these results demonstrate the superiority of the proposed approach compared to SOTA RGB trackers using deep features.

3) Parallel Analysis: Fig. 8 shows the parallel results of the top ten SOTAs with different features (i.e., hand-crafted features and deep features) on two types of videos (i.e., RGB videos and false-color videos). We find that trackers using deep features outperform those with hand-crafted features due to the discriminative and generalized capabilities derived from complex models and trained on sufficient data. Additionally, most trackers exhibit poorer performance on false-color videos than RGB videos. Notably, the AUC degradation of deep feature-based trackers is more significant than that of hand-crafted feature-based trackers, mainly due to the intrinsic differences in object representations between RGB and false-color videos, despite sharing three bands. The proposed SPIRIT tracker first leverages deep features learned from RGB and HS data for object representations, and then achieves intra and interband feature interaction, and finally is supported by a novel UM, achieving the optimal result. Overall, the proposed approach presents a promising solution for HS video object tracking, with potential applications in material discrimination and related fields.

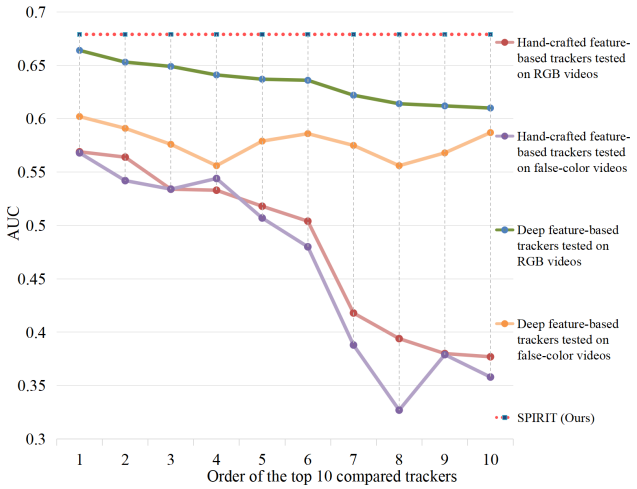


Fig. 8. Parallel comparisons with the top ten RGB trackers using hand-crafted and deep features. The top ten trackers are ranked by the AUC of RGB videos, For the hand-crafted feature-based trackers, the top ten trackers are STRCF, ARCF, AutoTrack, BACF, Staple, DSST, SAMF, DAT, CN, and KCF. For the deep feature-based trackers, the top ten trackers are SimTrack, SiamRPN++, SiamGAT, DiMP, Stark, SiamCAR, DaSiamRPN, ATOM, RTS, and SiamBAN.

C. Quantitative Evaluation With HS Trackers

In this section, we compare our SPIRIT with 12 representative HS trackers, including SEE-Net [7], SiamHYPER [8], MHT [11], MFI [13], TSCFW [15], BAHT [25], BAE-Net [27], SST-Net [28], SiamOHOT [50], CS_BACF [72], DeepHKCF [73], and convolutional network based hyperspectral tracking (CNHT) [74]. Table II summarizes the characteristics and experimental results of these trackers, sorted by AUCs. With an AUC of 0.679, SPIRIT exhibits competitive result and secures the overall first place, while CNHT performs the worst. SiamHYPER and SEE-Net achieved AUCs of 0.678 and 0.666, respectively, ranking the second and third places. In terms of the DP metric, SiamHYPER, SEE-Net, and SPIRIT produce 0.947, 0.934, and 0.925, respectively, leading to the top three places. From Table II, we observe that trackers utilizing deep features typically outperform those with hand-crafted features, as found in Fig. 8. For instance, the AUCs of SiamHYPER (0.678), SEE-Net (0.666),

TABLE II
DETAILS OF HS TRACKERS AND EXPERIMENTAL RESULTS

HS Trackers (Publications)	Frameworks	Features	UFB	TU	AUC	DP	FPS	MOP
CNHT (ICSM 2018)	KCF	Deep feature	Yes	Yes	0.170	0.336	2.6	CPU
DeepHKCF (TGRS 2019)	KCF	Deep feature	No	Yes	0.303	0.543	0.9	CPU
MHT (TIP 2019)	KCF	Hand-crafted feature	Yes	Yes	0.586	0.883	2.2	CPU
TSCFW (TGRS 2022)	KCF	Hand-crafted feature	Yes	Yes	0.587	0.870	3.4	CPU
CS_BACF (WHISPERS 2022)	KCF	Hand-crafted feature	No	Yes	0.589	0.897	11.5	CPU
MFI (WHISPERS 2021)	KCF	Hand-crafted feature+Deep feature	No	Yes	0.601	0.893	0.4	CPU
BAE-Net (ICIP 2020)	VITAL	Deep feature	Yes	Yes	0.606	0.879	0.5	GPU
SST-Net (WHISPERS 2021)	VITAL	Deep feature	Yes	Yes	0.623	0.917	0.5	GPU
BAHT (GRSL 2022)	SiamFC	Deep feature	No	No	0.632	0.905	16.0	GPU
SiamOHOT (TGRS 2023)	SiamFC	Deep feature	Yes	No	0.634	0.884	38.0	GPU
SEE-Net (TIP 2023)	SiamFC	Deep feature	Yes	No	0.666	0.934	8.7	GPU
SiamHYPER (TIP 2022)	SiamFC	Deep feature	Yes	No	0.678	0.947	19.0	GPU
SPIRIT (Ours)	SiamFC	Deep feature	Yes	Yes	0.679	0.925	26.0	GPU

The framework abbreviations used in the table are KCF, which stands for kernelized correlation filter, VITAL, which stands for visual tracking via adversarial learning, and SiamFC, which represents fully convolutional Siamese network. UFB denotes the attempt to use of full-band. TU denotes template update, and MOP represents the main operation platform.

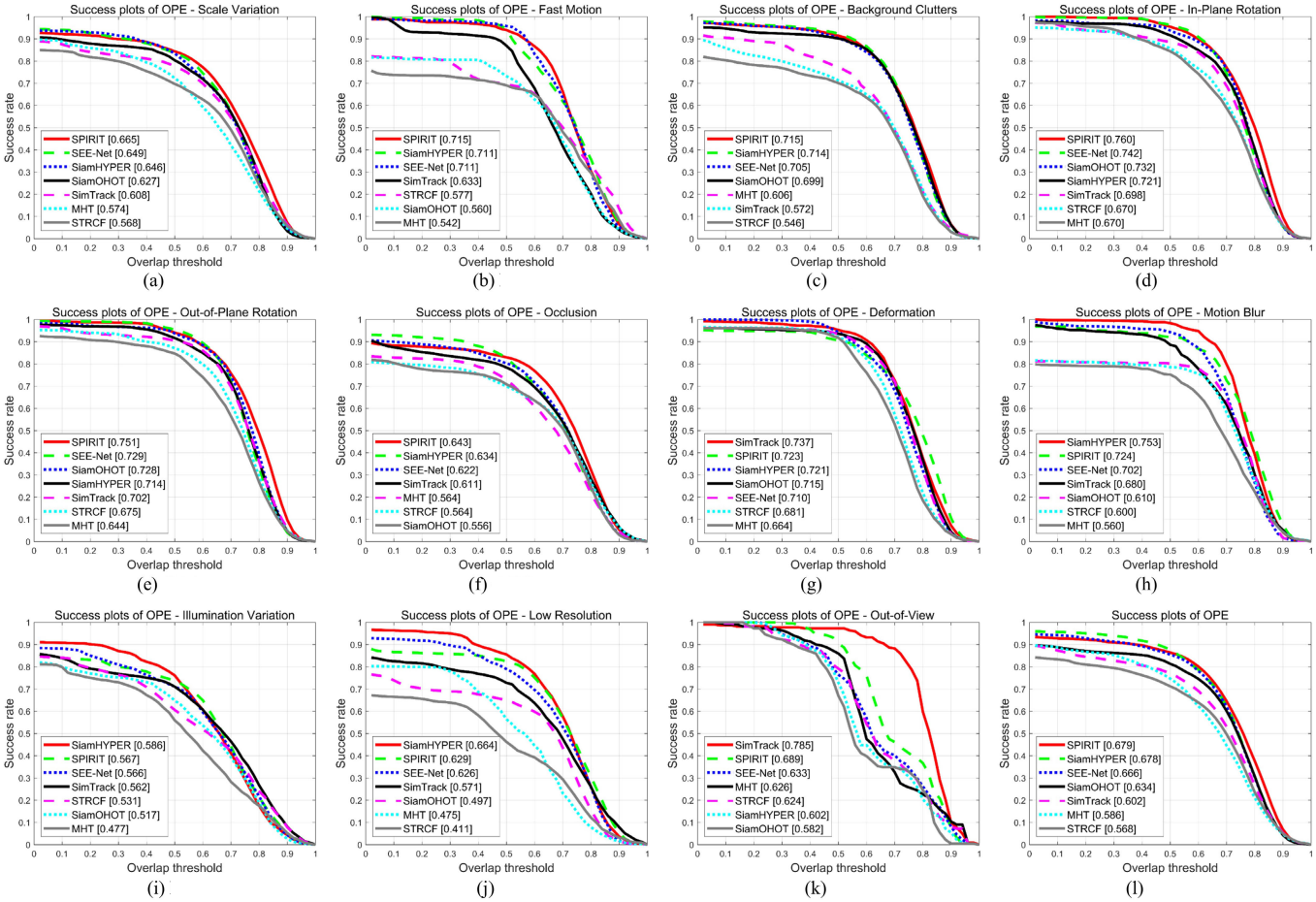


Fig. 9. Success plots for (a)–(k) individual attributes and (l) overall. Six SOTA trackers are compared with the proposed SPIRIT. (a) SV. (b) FM. (c) BC. (d) IPR. (e) OPR. (f) OCC. (g) DEF. (h) MB. (i) IV. (j) LR. (k) OV. (l) overall.

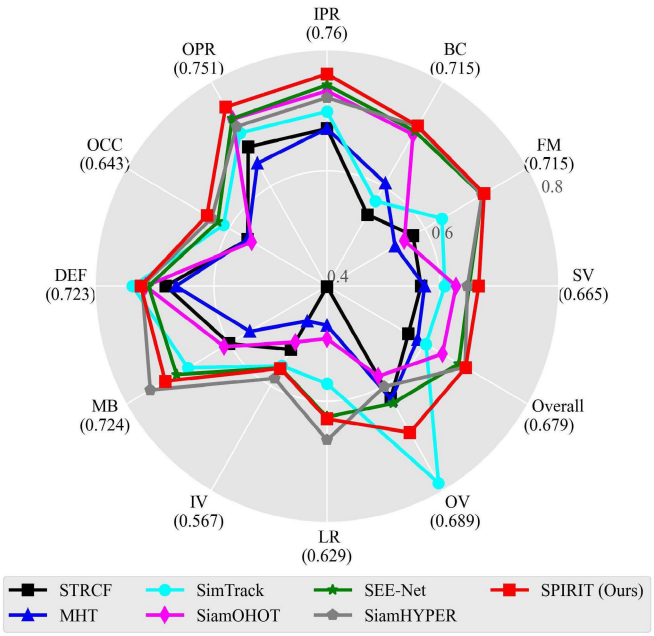


Fig. 10. Radar plot for individual attributes and overall in terms of AUC. Values in parentheses are AUCs of the proposed SPIRIT tracker.

and SiamOHOT (0.634) surpass those of CS_BACF (0.589), TSCFW (0.587), and MHT (0.586). Furthermore, we observe that Siamese-based trackers generally outperform those with correlation filters, as Siamese networks enable the learning of discriminative features. Although CNHT and DeepHKCF utilize deep features for HS object tracking, the KCF framework intrinsically limits their performance. The proposed SPIRIT tracker inherits the Siamese tracker and can leverage deep features to acquire discriminative features.

It is noteworthy that recent SOTAs have made efforts to explore full-band information for HS object tracking, such as SiamHYPER, SEE-Net, SiamOHOT, SST-Net, BAE-Net, and TSCFW. These methods are promising as the utilization of full-band information enables trackers to potentially become material-aware and acquire more discriminative cues, thereby yielding robust performance. In addition, MFI, TSCFW, DeepHKCF, and CNHT can update their models to adapt to object changes, but the hand-crafted features or model structure would limit their performance. Meanwhile, BAE-Net and SST-Net are online trackers with acceptable accuracy but low speed. For Siamese-based HS trackers, such as SiamHYPER, SEE-Net, and SiamOHOT, the templates are typically fixed in the first frame, which may result in tracking drift in case of significant object changes. Overall, the proposed SPIRIT method inherits the Siamese framework and endeavors to acquire full-band information. Moreover, SPIRIT also contains a UM to deal with object changes for producing competitive results.

D. Attribute-Based Evaluation

To test the properties of trackers, we further conduct experiments on 11 challenging attributes. For simplicity, we only report the results of excellent RGB trackers (i.e., SimTrack [4] and STRCF [62]) and HS trackers (i.e., SEE-Net [7], SiamHYPER [8], MHT [11], and SiamOHOT [50]). SimTrack and STRCF are run on false-color videos, while the rest are tested on HS videos. Figs. 9 and 10 exhibit the success and radar plots for individual attributes and overall, respectively. It is observed that SPIRIT ranks first in six (i.e., SV, FM, BC, IPR, OPR, and OCC) out of 11 attributes. BC is the most prevalent attribute. For this attribute, the AUCs of SPIRIT (0.715), SiamHYPER (0.714), SEE-Net (0.705), SiamOHOT (0.699), and MHT (0.606) are significantly superior to those of RGB trackers, due to exploiting rich spectral information. SV is another challenging attribute, and SPIRIT achieves the optimal AUC of 0.665, which is 1.6% and 1.9% higher than the AUCs of the second-ranked SEE-Net (0.649) and the third-ranked SiamHYPER (0.646), respectively. In the remaining attributes including DEF, MB, IV, LR, and OV, SPIRIT ranks the second place, and its performance is comparable or better than that of SiamHYPER, SEE-Net, and SimTrack. Noticed that SimTrack shows a significant boost over SPIRIT in the OV attribute. This is because SimTrack jointly implements the feature extraction and interaction by a transformer backbone, which helps to remove elaborate IMs and consider invariant part-level cues for tracking. In general, experimental results demonstrate that SPIRIT can handle challenging attributes and achieve the optimal overall AUC of 0.679.

E. Running Speed Comparison

We compared the running speeds of several SOTA HS trackers, as shown in Table II and Fig. 1. These trackers including SST-Net (0.5 frames/s), BAE-Net (0.5 frames/s), MFI (0.4 frames/s), and DeepHKCF (0.9 frames/s) have relatively slower tracking speeds than SiamHYPER (19.0 frames/s), SEE-Net (8.7 frames/s), BAHT (16.0 frames/s), and CS_BACF (11.5 frames/s). Only the proposed SPIRIT and SiamOHOT achieve high processing speeds of 26.0 and 38.0 frames/s, respectively. In addition, the proposed SPIRIT achieves superior tracking accuracy than SiamOHOT. It is found that SEE-Net, SiamOHOT, and SPIRIT all inherit the framework of SiamFC [52], but there are significant differences in the running speeds. This is because SiamOHOT focuses more on using knowledge distillation techniques to refine the model to improve tracking efficiency. While SEE-Net incorporates the decision-level fusion strategy to improve tracking effectiveness but introduces an additional computational burden. Our approach adopts a feature-level fusion strategy to ensure tracking efficiency and introduces the SAM, IM, and UM to improve tracking effectiveness. HS videos are typically captured at a frame rate of 25 frames/s. Therefore, the proposed SPIRIT can process HS videos in real-time while maintaining competitive tracking accuracy, making it suitable for on-board HS data processing.

F. Visual Comparison

In this section, we conduct a qualitative comparison with six SOTAs including SimTrack [4], SEE-Net [7], SiamHYPER [8], MHT [11], SiamOHOT [50], and STRCF [62]. These HS trackers are run on HS videos, while SimTrack and STRCF are tested on false-color videos. Fig. 11 presents visual

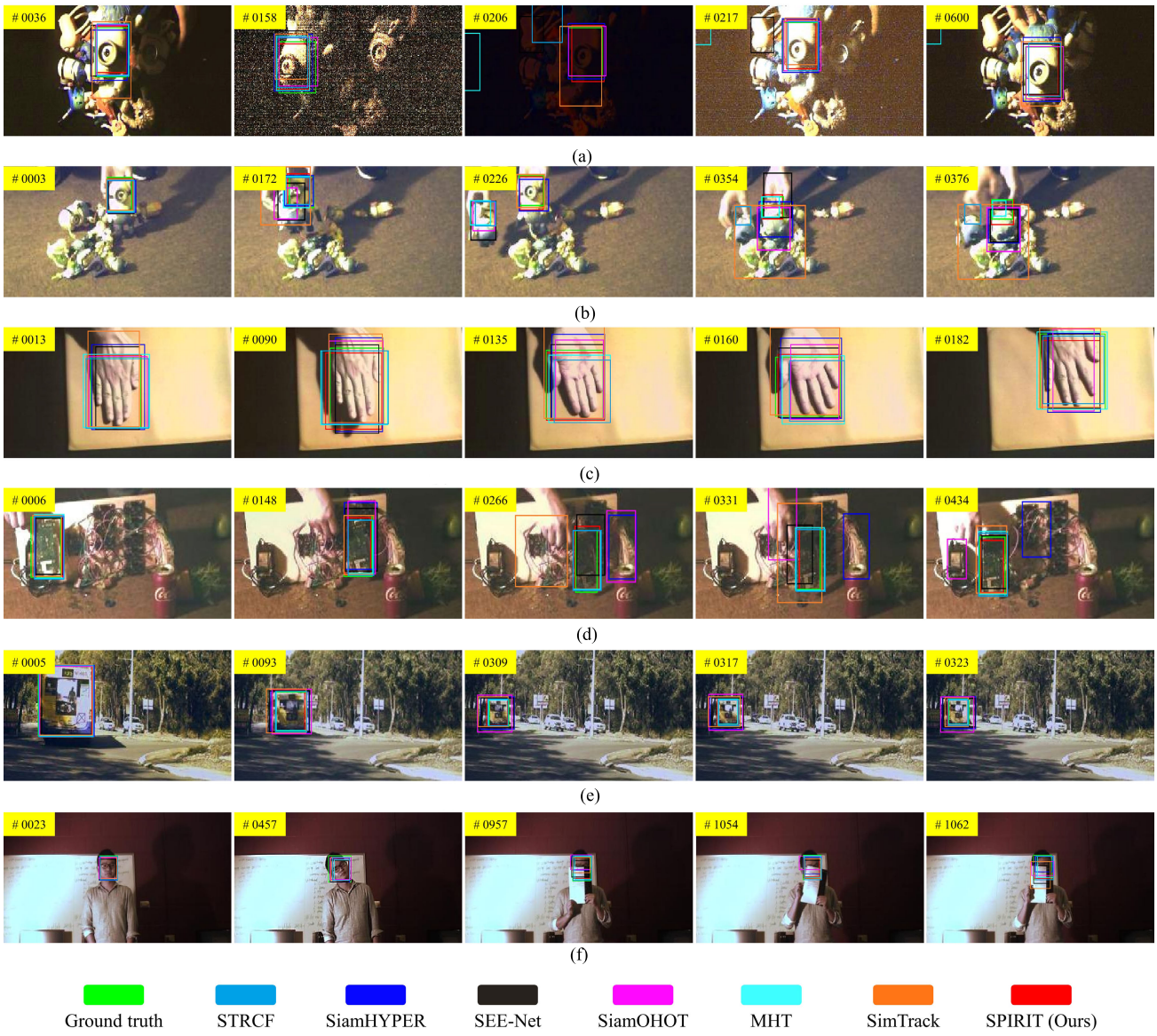


Fig. 11. Qualitative comparisons of the proposed SPIRIT tracker with six SOTA trackers. Results are shown in corresponding false-color images, with the current frame displayed in the upper-left corner of each image. (a) *toy*2, attribute: BC, OCC, SV, IV, and OPR. (b) *toy*1, attribute: BC and OCC. (c) *hand*, attribute: BC, SV, DEF, and OPR. (d) *board*, attribute: IPR, OPR, BC, OCC, and SV. (e) *bus*2, attribute: IV, SV, OCC, and FM. (f) *face*2, attribute: IPR, OPR, SV, and OCC.

examples of six false-color videos, including *toy*2, *toy*1, *hand*, *board*, *bus*2, and *face*2. These videos cover diverse challenging attributes. In the *toy*2 video, the object undergoes BC, OCC, SV, IV, and OPR. It is noted that SimTrack inaccurately estimates the object scale at frames #0036 and #0206, while MHT and STRCF lose the object at frames #0206 and #0217 due to significant illuminance variation. In the *board* video, SiamHYPE, SEE-Net, SiamOHOT, and SimTrack encounter varying degrees of tracking drifts, while SPIRIT is able to track the object effectively. In other cases shown in Fig. 11, SPIRIT demonstrates better performance in handling challenging attributes. Overall, the qualitative evaluation highlights the robustness of the proposed SPIRIT approach, making it a promising candidate for HS video object tracking applications.

G. Ablation Study

In this section, we conduct a series of ablation experiments to validate the major components including the SAM, IM, and UM. For this purpose, five variants including the Baseline, Variant_1, Variant_2, Variant_3, and Variant_4 are constructed and compared with the SPIRIT tracker on HS videos. Table III summarizes the components and experimental results, while Fig. 12 presents the success plot and precision plot of these variants and the SPIRIT tracker. The Baseline serves as the baseline approach, which adopts the band grouping strategy of Sequential and the feature interaction of Addition to realize HS object tracking without dynamically updating the template. Variant_1 integrates the proposed SAM on top of the Baseline for band grouping. On the basis of Variant_1, Variant_2 introduces the interband feature interaction of the proposed IM. Variant_3 further integrates the intraband feature interaction of the IM on top of Variant_2. Moreover, Variant_4 indicates the removal of the intraband feature interaction from the proposed SPIRIT method that integrates components, including SAM, interband and IBI of IM, and UM.

TABLE III
DETAILS OF VARIANTS AND EXPERIMENTAL RESULTS

Trackers	Band grouping strategy		Feature interaction			Template update	AUC	AUC↑	DP	DP↑
	Sequential	SAM (Ours)	Addition	IM (Ours)		UM (Ours)				
				Inter-band	Intra-band					
Baseline	Yes	No	Yes	No	No	No	0.418	n/a	0.618	n/a
Variant_1	No	Yes	Yes	No	No	No	0.593	17.5%	0.812	19.4%
Variant_2	No	Yes	No	Yes	No	No	0.628	21.0%	0.845	22.7%
Variant_3	No	Yes	No	Yes	Yes	No	0.640	22.2%	0.865	24.7%
Variant_4	No	Yes	No	Yes	No	Yes	0.650	23.2%	0.884	26.6%
SPIRIT	No	Yes	No	Yes	Yes	Yes	0.679	26.1%	0.925	30.7%

SAM, IM, and UM indicate the spectral awareness module, interaction module, and update module, respectively, in the proposed method. The Sequential approach groups adjacent three bands of a HS image to generate false-color images. Addition approach sums up the features generated from false-color images to achieve feature interaction. AUC↑ and DP↑ denote the improvement of the current tracker's AUC and DP compared to the Baseline, respectively..

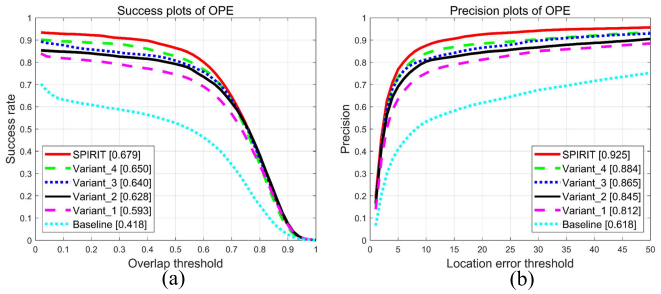


Fig. 12. (a) Success plot and (b) precision plot of the variants tested on HS videos.

As shown in Table III, the Baseline attains an AUC of 0.418 and a DP of 0.618. However, the introduction of the SAM in Variant_1 results in a significant improvement in both AUC and DP. Specifically, there is a 17.5% increase in AUC (from 0.418 to 0.593) and a 19.4% increase in DP (from 0.618 to 0.812). This can be attributed to the ability of SAM to divide the HS image into false-color images with low correlation, thereby reducing information redundancy that may exist in a Sequential fashion. Additionally, the SAM generates the contributions of false-color images that can be utilized to facilitate interband interactions of the IM. By comparing Variant_2 to Variant_1, it is evident that the AUC and DP of Variant_2 are improved by 3.5% and 3.3%, respectively. This implies that the SAM not only produces low correlation false-color images but also enhances interband feature interactions of the IM, leading to an improvement in performance. However, Variant_2 only considers interband interaction, with the AUC and DP of 0.628 and 0.845, respectively, while ignoring the potential for exploiting intraband feature interactions. Upon comparing Variant_2 and Variant_3, we find that the addition of IBI in Variant_3 leads to 1.2% and 2.0% increases in AUC and DP, respectively. This is because IBI, built on the foundation of interband interaction, enables Variant_3 to learn more robust object representations, which in turn facilitates HS object tracking. Similar conclusions can be drawn by comparing Variant_4 with SPIRIT. The IBI network improves performance at the cost of only a few computational burdens, achieving a trade-off between efficiency and effectiveness. Compared to Variant_1, Variant_3 shows a 4.7% increase in AUC and a 5.3% increase in DP, further validating the effectiveness of the proposed IM. To prove the role of the UM, we conduct experimental comparisons with and without dynamically updated templates. A comparison between SPIRIT and Variant_3 reveals that dynamically updating the template increases the AUC and DP by 3.9% and 6.0%, respectively. With the aid of the UM, the proposed SPIRIT is capable of acquiring spectral, spatial, and temporal information to adapt to object changes in appearance, resulting in satisfactory tracking results. Similar conclusions can be drawn by comparing Variant_4 and Variant_2. In summary, experimental results validate the effectiveness of major components of the proposed method. As a result, the proposed SPIRIT is able to produce competitive performance through comprehensive utilization of these components.

V. CONCLUSION

This article presents an end-to-end deep learning network SPIRIT for HS video object tracking. First, the proposed method constructs a SAM that evaluates the band contributions to downstream missions by considering nonlinear and global interactions from the perspective of spectral reconstruction. Guided by this evaluation, the HS image is adaptively divided into multiple false-color images, which are fed to a feature extraction module in a transferred tracking network pretrained with RGB data to obtain robust object representations by exploiting the full-band information. Subsequently, an IM is proposed to achieve interband and intraband feature interaction while ensuring tracking speed. Furthermore, the proposed SPIRIT contains a novel UM that evaluates the tracking confidence to adapt to object changes and attenuate tracking drifts. Extensive experiments demonstrate that the proposed method achieves the optimal trade-off between effectiveness and efficiency.

Nevertheless, our SPIRIT has some shortcomings. First, the feature extraction process consumes a lot of computational cost. Second, the motion information contained in the video

frames is ignored, leading to limited adaptability to complex scenes. Therefore, more attention should be paid to simplifying the model structure. Besides, using transformer attention to synergize the appearance and motion information of the object may also be a promising solution.

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. Javed, M. Danelljan, F. S. Khan, M. H. Khan, M. Felsberg, and J. Matas, "Visual object tracking with discriminative filters and Siamese networks: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 6552–6574, May 2023.

[3] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11920–11929.

[4] B. Chen et al., "Backbone is all your need: A simplified architecture for visual object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 375–392.

[5] 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.

[6] 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. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 6687–6698, 2022.

[7] 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.

[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] J. He, J. Li, Q. Yuan, H. Shen, and L. Zhang, "Spectral response function-guided deep optimization-driven network for spectral super-resolution," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4213–4227, Sep. 2022.

[10] J. He et al., "Spectral super-resolution meets deep learning: Achievements and challenges," *Inf. Fusion*, vol. 97, Sep. 2023, Art. no. 101812.

[11] F. Xiong, J. Zhou, and Y. Qian, "Material based object tracking in hyperspectral videos," *IEEE Trans. Image Process.*, vol. 29, pp. 3719–3733, 2020.

[12] Z. Liu et al., "An anchor-free Siamese target tracking network for hyperspectral video," in *Proc. 11th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Mar. 2021, pp. 1–5.

[13] Z. Zhang, K. Qian, J. Du, and H. Zhou, "Multi-features integration based hyperspectral videos tracker," in *Proc. 11th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Mar. 2021, pp. 1–5.

[14] 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 Understand.*, vol. 223, Oct. 2022, Art. no. 103535.

[15] 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.*, vol. 60, 2022, Art. no. 5541012.

[16] Y. Xiao, X. Su, Q. Yuan, D. Liu, H. Shen, and L. Zhang, "Satellite video super-resolution via multiscale deformable convolution alignment and temporal grouping projection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–19, 2022, Art. no. 5610819, doi: 10.1109/TGRS.2021.3107352.

[17] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4277–4286.

[18] Y. Xiao, Q. Yuan, K. Jiang, J. He, Y. Wang, and L. Zhang, "From degrade to upgrade: Learning a self-supervised degradation guided adaptive network for blind remote sensing image super-resolution," *Inf. Fusion*, vol. 96, pp. 297–311, Aug. 2023.

[19] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: Accurate tracking by overlap maximization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4655–4664.

[20] Y. Xiao et al., "Local–global temporal difference learning for satellite video super-resolution," *IEEE Trans. Circuits Syst. Video Technol.*, 2023, doi: 10.1109/TCSVT.2023.3312321.

[21] L. Huang, X. Zhao, and K. 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.

[22] M. Müller, A. Bibi, S. Giancola, S. Alsubaihi, and B. Ghanem, "TrackingNet: A large-scale dataset and benchmark for object tracking in the wild," in *Proc. 15th Eur. Conf. Comput. Vis. (ECCV)*, vol. 11205. Munich, Germany, 2018, pp. 310–327.

[23] H. Fan et al., "LaSOT: A high-quality benchmark for large-scale single object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5369–5378.

[24] Y. Su, S. Mei, G. Zhang, Y. Wang, M. He, and Q. Du, "Gaussian information entropy based band reduction for unsupervised hyperspectral video tracking," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 791–794.

[25] 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.*, vol. 19, pp. 1–5, 2022.

[26] S. Wang, K. Qian, and P. Chen, "BS-SiamRPN: Hyperspectral video tracking based on band selection and the Siamese region proposal network," in *Proc. 12th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens.*, Sep. 2022, pp. 1–8.

[27] 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 *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 2106–2110.

[28] Z. Li, X. Ye, F. Xiong, J. Lu, J. Zhou, and Y. Qian, "Spectral–spatial–temporal attention network for hyperspectral tracking," in *Proc. 11th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Mar. 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.*, vol. 19, pp. 1–5, 2022.

[30] Z. Li, F. Xiong, J. Lu, J. Zhou, and Y. Qian, "Material-guided Siamese fusion network for hyperspectral object tracking," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2809–2813.

[31] Y. Wang, Y. Liu, G. Zhang, Y. Su, S. Zhang, and S. Mei, "Spectral–spatial-aware transformer fusion network for hyperspectral object tracking," in *Proc. 12th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Sep. 2022, pp. 1–5.

[32] S. You, H. Zhu, M. Li, and Y. Li, "A review of visual trackers and analysis of its application to mobile robot," 2019, *arXiv:1910.09761*.

[33] B. Han, D. Comaniciu, Z. Ying, and L. S. Davis, "Sequential kernel density approximation and its application to real-time visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1186–1197, Jul. 2008.

[34] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T. S. Huang, and S. Yan, "Sparse representation for computer vision and pattern recognition," *Proc. IEEE*, vol. 98, no. 6, pp. 1031–1044, Jun. 2010.

[35] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1296–1311, Oct. 2003.

[36] D. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[37] 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. 583–596, Mar. 2015.

[38] M. Danelljan, G. Häger, 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.

[39] 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)*, Jun. 2016, pp. 1401–1409.

[40] Y. Chen, Y. Tang, T. Han, Y. Zhang, B. Zou, and H. Feng, "RAMC: A rotation adaptive tracker with motion constraint for satellite video single-object tracking," *Remote Sens.*, vol. 14, no. 13, p. 3108, Jun. 2022.

[41] 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.*, Jun. 2014, pp. 1090–1097.

[42] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[43] S. Gao, C. Zhou, C. Ma, X. Wang, and J. Yuan, "AiATrack: Attention in attention for transformer visual tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 146–164.
[44] L. Lin, H. Fan, Z. Zhang, Y. Xu, and H. Ling, "SwinTrack: A simple and strong baseline for transformer tracking," in *Proc. NeurIPS*, 2022, pp. 16743–16754.
[45] J. He, Q. Yuan, J. Li, and L. Zhang, "PoNet: A universal physical optimization-based spectral super-resolution network for arbitrary multispectral images," *Inf. Fusion*, vol. 80, pp. 205–225, Apr. 2022.
[46] J. He, Q. Yuan, J. Li, Y. Xiao, and L. Zhang, "A self-supervised remote sensing image fusion framework with dual-stage self-learning and spectral super-resolution injection," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 131–144, Oct. 2023.
[47] A. Banerjee, P. Burlina, and J. Broadwater, "Hyperspectral video for illumination-invariant tracking," in *Proc. 1st Workshop Hyperspectral Image Signal Process., Evol. Remote Sens.*, Aug. 2009, pp. 474–477.
[48] H. Van Nguyen, A. Banerjee, and R. Chellappa, "Tracking via object reflectance using a hyperspectral video camera," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 44–51.
[49] R. T. Collins, "Mean-shift blob tracking through scale space," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jul. 2003, p. 234.
[50] C. Sun, X. Wang, Z. Liu, Y. Wan, L. Zhang, and Y. Zhong, "SiamOHOT: A lightweight dual Siamese network for onboard hyperspectral object tracking via joint spatial–spectral knowledge distillation," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023, Art. no. 5521112, doi: 10.1109/TGRS.2023.3307052.
[51] 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, Art. no. 5506114, doi: 10.1109/TGRS.2023.3253173.
[52] Y. Tang, H. Huang, Y. Liu, and Y. Li, "A Siamese network-based tracking framework for hyperspectral video," *Neural Comput. Appl.*, vol. 35, no. 3, pp. 2381–2397, Jan. 2023.
[53] Y. Cai, X. Liu, and Z. Cai, "BS-Nets: An end-to-end framework for band selection of hyperspectral image," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1969–1984, Mar. 2020.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2016, pp. 770–778.
[55] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 10428–10437.
[56] Z. H. Zheng et al., "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. 34th AAAI Conf. Artif. Intell./32nd Innov. Appl. Artif. Intell. Conf./10th AAAI Symp. Educ. Adv. Artif. Intell.*, vol. 34. New York, NY, USA, 2020, pp. 12993–13000.
[57] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[58] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2017, pp. 1–19.
[59] Y. Wu, J. Lim, and M.-H. Yang, "Object tracking benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1834–1848, Sep. 2015.
[60] Z. Huang, C. Fu, Y. Li, F. Lin, and P. Lu, "Learning aberrance repressed correlation filters for real-time UAV tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2891–2900.
[61] H. K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1144–1152.
[62] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang, "Learning spatial–temporal regularized correlation filters for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4904–4913.
[63] Z. Liu, Z. Lian, and Y. Li, "A novel adaptive kernel correlation filter tracker with multiple feature integration," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 254–265.
[64] H. Possegger, T. Mauthner, and H. Bischof, "In defense of color-based model-free tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2113–2120.
[65] 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)*, Jul. 2017, pp. 6931–6939.
[66] Z. Chen, B. Zhong, G. Li, S. Zhang, and R. Ji, "Siamese box adaptive network for visual tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6667–6676.
[67] 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/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 15175–15184.
[68] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware Siamese networks for visual object tracking," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 103–119.
[69] G. Bhat, M. Danelljan, L. Van Gool, and R. Timofte, "Learning discriminative model prediction for tracking," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6181–6190.
[70] M. Paul, M. Danelljan, C. Mayer, and L. Van Gool, "Robust visual tracking by segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 571–588.
[71] D. Guo, Y. Shao, Y. Cui, Z. Wang, L. Zhang, and C. Shen, "Graph attention tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9538–9547.
[72] Y. Zhang, X. Li, F. Wang, B. Wei, and L. Li, "A fast hyperspectral object tracking method based on channel selection strategy," in *Proc. 12th Workshop Hyperspectral Imag. Signal Process., Evol. Remote Sens. (WHISPERS)*, Sep. 2022, pp. 1–5.
[73] B. Uzkent, A. Rangnekar, and M. J. Hoffman, "Tracking in aerial hyperspectral videos using deep kernelized correlation filters," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 449–461, Jan. 2019.
[74] K. Qian, J. Zhou, F. Xiong, H. Zhou, and J. Du, "Object tracking in hyperspectral videos with convolutional features and kernelized correlation filter," in *Proc. Int. Conf. Smart Multimedia*, 2018, pp. 308–319.

Yuzeng Chen received the B.S. degree in geographic information science from the Southwest University of Science and Technology, Mianyang, China, in 2020, and the M.S. degree from Central South University, Changsha, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China.

His research interests include remote sensing/hyperspectral video object detection and tracking.

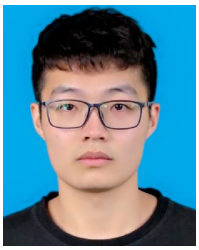
Qiangqiang Yuan (Member, IEEE) received the B.S. degree in surveying and mapping engineering and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2006 and 2012, respectively.

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

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

Yuqi Tang (Member, IEEE) received the Ph.D. degree in photogrammetric and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2013.

Since 2017, she has been an Associate Professor with the School of Geosciences and Info-Physics, Central South University, Changsha, China. Her research interests include object identification/tracking, land-cover/-use classification and change detection in multisource remote sensing images, and natural resource monitoring.



Yi Xiao (Graduate Student Member, IEEE) received the B.S. degree from the School of Mathematics and Physics, China University of Geosciences, Wuhan, China, in 2020, and the M.S. degree from Wuhan University, Wuhan, in 2022, where he is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics.

His major research interests include remote sensing image/video processing and computer vision. More details can be found at https://xy-boy.github.io.

Jiang He (Graduate Student Member, IEEE) received the B.S. degree in remote sensing science and technology from the Faculty of Geosciences and Environmental Engineering, Southwest Jiaotong University, Chengdu, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Geodesy and Geomatics, Wuhan University, Wuhan, China.

His research interests include hyperspectral super-resolution, image fusion, quality improvement, remote sensing image processing, and deep learning.

Liangpei Zhang (Fellow, IEEE) received the B.S. degree in physics from Hunan Normal University, Changsha, China, in 1982, the M.S. degree in optics from the Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 1998.

He was a Principal Scientist of the China State Key Basic Research Project from 2011 to 2016 appointed by the Ministry of National Science and Technology of China to lead the Remote Sensing Program in China. He is currently a "Chang-Jiang Scholar" Chair Professor appointed by the Ministry of Education of China with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing (LIESMARS), Wuhan University. He has published more than 700 research articles and five books. He is the Institute for Scientific Information (ISI) Highly Cited Author. He holds 30 patents. His research interests include hyperspectral remote sensing, high-resolution remote sensing, image processing, and artificial intelligence.

Dr. Zhang is a fellow of the Institution of Engineering and Technology (IET). He was a recipient of the 2010 Best Paper Boeing Award, the 2013 Best Paper ERDAS Award from the American Society of Photogrammetry and Remote Sensing (ASPRS), and the 2016 Best Paper Theoretical Innovation Award from the International Society for Optics and Photonics (SPIE). His research teams won the top three prizes in the IEEE Geoscience and Remote Sensing Society (GRSS) 2014 Data Fusion Contest. His students have been selected as the Winners or a Finalists of the IEEE International Geoscience and Remote Sensing Symposium (IGARSS) Student Paper Contest in recent years. He is also the Founding Chair of the IEEE GRSS Wuhan Chapter. He also serves as an associate editor or an editor for more than ten international journals. He is also serving as an Associate Editor for the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING.