Article:
Gao, L., Chen, L., Jiang, Y. et al. (2025) Feature-level fusion network for hyperspectral object tracking via mixed multi-head self-attention learning. Remote Sensing, 17 (6). 997. ISSN: 2072-4292

https://doi.org/10.3390/rs17060997

remote sensing

MDPI

Article

Feature-Level Fusion Network for Hyperspectral Object Tracking via Mixed Multi-Head Self-Attention Learning

Long Gao [1], Langkun Chen [1], Yan Jiang [2], Bobo Xi [1], Weiying Xie [1,*] and Yunsong Li [1]

[1] The State Key Laboratory of Integrated Service Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China; lgao@xidian.edu.cn (L.G.); 22011211080@stu.xidian.edu.cn (L.C.); xibobo@xidian.edu.cn (B.X.); ysli@mail.xidian.edu.cn (Y.L.)

[2] The Department of Electronic and Electrical Engineering, The University of Sheffield, Sheffield S10 2TN, UK; yjiang71@sheffield.ac.uk

* Correspondence: wyxie@xidian.edu.cn

Abstract: Hyperspectral object tracking has emerged as a promising task in visual object tracking. The rich spectral information within hyperspectral images benefits the accurate tracking in challenging scenarios. The performances of existing hyperspectral object tracking networks are constrained by neglecting the interactive information among bands within hyperspectral images. Moreover, designing an accurate deep learning-based algorithm for hyperspectral object tracking poses challenges because of the substantial amount of training data required. In order to address these challenges, a new mixed multi-head attention-based feature fusion tracking (MMFT) algorithm for hyperspectral videos is proposed. Firstly, MMFT introduces a feature-level fusion module, mixed multi-head attention feature fusion (MMFF), which fuses false-color features and augments the fused feature with one mixed multi-head attention (MMA) block with interactive information, which increases the representational ability of the features for tracking. Specifically, MMA learns the interactive information across the bands in the false-color images and incorporates the learned interactive information into the fused feature, which is obtained by combining the features of the false-color images. Secondly, a new training procedure is introduced, in which the modules designed for hyperspectral object tracking are first pre-trained on a sufficient amount of modified RGB data to enhance generalization, and then fine-tuned on a limited amount of HS data for task adaption. Extensive experiments verify the effectiveness of MMFT, demonstrating its SOTA performance.

Keywords: feature fusion; mixed multi-head attention; Transformer; hyperspectral object tracking

Academic Editor: Javier Marcello

Received: 19 January 2025
Revised: 10 March 2025
Accepted: 11 March 2025
Published: 12 March 2025

Citation: Gao, L.; Chen, L.; Jiang, Y.; Xi, B.; Xie, W.; Li, Y. Feature-Level Fusion Network for Hyperspectral Object Tracking via Mixed Multi-Head Self-Attention Learning. *Remote Sens.* **2025**, *17*, 997. https://doi.org/10.3390/rs17060997

1. Introduction

Hyperspectral object tracking has attracted increasing attention, since the material information contained in hyperspectral (HS) images enhances a tracker's discriminative ability against background clutter [1,2]. Unlike RGB images, which capture only three channels, HS images span a wide range of wavelengths. This enables the differentiation of objects that appear similar in RGB but exhibit distinct spectral characteristics, thereby enhancing tracking performance in challenging scenarios. However, the datasets for HS object tracking are small, which makes designing a robust tracker for HS object tracking challenging. There has been a concerted focus on exploiting deep learning and traditional machine learning methods to track targets with spectral information [3–5]. By transferring

the discriminative capabilities acquired from extensive RGB datasets, deep learning-driven trackers significantly outperform traditional machine learning methods [6–8].

Deep learning-based HS object tracking methods typically develop their networks by using RGB tracking networks and applying band regrouping or selection methods to convert HS data into false-color data to fit the network [9–12]. Some works reduce the number of channels of the HS images to one three-channel false-color image and process the images with RGB tracking networks [9,13]. These trackers achieve high speeds, e.g., SSDT-Net, which runs at 36 frames per second (FPS) [13] and processes only one three-channel image at a time for tracking within the current frame. Other works divide HS images into multiple false-color images and track the target by implementing several RGB tracking networks [5,6,14]. Specifically, a sixteen-channel HS image is segmented into five distinct false-color representations, and the tracking result is derived by combining the tracking results of the target in each of these false-color images separately. Compared with methods based on a single false-color image, tracking a target with multiple false-color images shows better performance as more spectral information is preserved. However, band regrouping methods neglect interactive information across the bands in HS images, limiting performance improvement.

Additionally, HS object tracking methods face the challenge of limited HS data. Deep learning-based methods, particularly those based on Transformer, require substantial labeled training data to grasp long-range dependencies. The scarcity of available HS data often results in overfitting. This issue is compounded by the fact that generating labeled HS datasets is expensive and time-consuming, making it difficult to train robust tracking networks capable of handling the complex characteristics of HS videos.

In order to tackle these challenges, a novel HS tracker, the mixed multi-head attention-based feature fusion tracking (MMFT) algorithm, is proposed. Firstly, mixed multi-head attention feature fusion (MMFF) is proposed to combine false-color features and learn the interactive information between bands in one mixed multi-head attention (MMA) block. MMFF obtains a fused feature by adding the false-color features using learnable weights, which do not contain the interactive information. Simply conducting multi-head self-attention (MHSA) on the fused feature to learn the interactive information is insufficient since MHSA learns only the interactive information within the fused feature. Directly conducting MHSA on the false-color images to learn interactive information requires performing MHSA on multiple features, which is inefficient. Therefore, MMA divides the heads of MHSA learning into two groups. One group is for self-attention learning on the fused feature, while the other is for cross-attention learning on the fused feature and the false-color features. This methodology enables learning of the interactive information from the mixed spectral information and the false-color images. The learned interactive information is also integrated into the fused feature by MMFF, which enhances the feature's representational capacity for tracking tasks. Additionally, compared to conducting MHSA on the false-color images, MMFF is more efficient since the queries in the self-attention and the cross-attention learning are formed from the fused feature, which contains the same number of tokens as one feature of a false-color image. Secondly, a two-step (TS) training strategy is proposed to mitigate overfitting resulting from the limited amount of HS data. Training MMFF or fine-tuning the whole MMFT network on HS data cannot achieve satisfactory performance due to overfitting. In the new training procedure, the MMFF module is pre-trained with modified RGB data from a large RGB tracking dataset to enhance generalization, while the rest of the algorithm is frozen. Subsequently, MMFT performs fine-tuning on the HS dataset to adjust its parameters specifically for the HS object tracking task. The experimental results validate that the new training procedure enhances

the tracker's performance. Extensive experiments on the HS object tracking dataset verify that MMFT achieves SOTA performance.

As shown in Figure 1, MMFT demonstrates a superior AUC score compared to other HS trackers, achieving a tracking speed of 26.1 FPS. Although SSDT-Net operates at a faster speed than MMFT, its tracking performance is considerably lower.

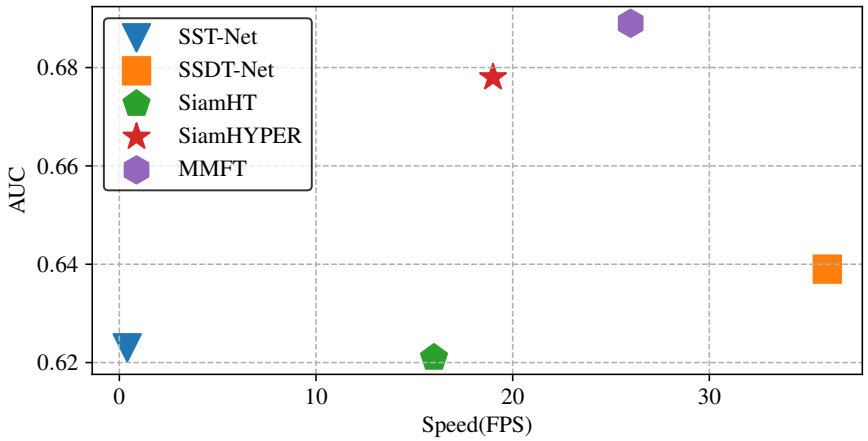

Figure 1. A comparison of FPS and AUC between MMFT and other HS trackers.

In general, the paper's primary contributions can be summarized as follows:

1. A novel feature fusion module, MMFF, is proposed, which enhances the hyperspectral object tracking network by integrating interactive information between spectral bands through an MMA mechanism.
2. A two-step (TS) training procedure is introduced, which effectively addresses the challenge of limited HS data by pre-training with large RGB datasets and fine-tuning on a smaller HS dataset, improving generalization and task adaptation.
3. Based on the proposed methods, a new HS tracking method, MMFT, is introduced. Comprehensive experiments on the dataset for HS object tracking confirm the effectiveness of MMFT, highlighting its impressive performance with real-time inference speed.

The structure of this paper is as follows. Section 2 provides a brief review of related methods. Section 3 presents the details of MMFT. Section 4 shows the experimental results with a comprehensive analysis. Section 5 concludes this work.

2. Related Work

An overview of research relevant to this study is presented in this section. Object tracking algorithms based on RGB data are introduced first. Then, works related to HS object tracking are reviewed. Moreover, methods based on Transformer in object tracking are introduced.

2.1. RGB Object Tracking

The goal of visual object tracking is to estimate the location and bounding box of a specified object within a video sequence. Most efforts have been devoted to tracking methods on RGB videos. The Siamese network framework emerged as a prevalent architecture in object tracking, following the introduction of SiamFC [15]. Siamese-based methods formulated object tracking as a matching problem, aiming to locate samples resembling the template within search patches. The template patch was obtained from the first frame of a video, while the search patches were acquired from the subsequent frames. Numerous advancements were explored to advance the precision of Siamese-based

tracking algorithms [16–19]. For instance, SiamRPN [16] and SiamRPN++ [18] integrated the Region Proposal Network [20] into object tracking, and estimated the states of the target through dual branches dedicated to classification and regression. Anchor-free methods were also integrated into Siamese-based tracking networks to alleviate dependencies on hyperparameters [19,21,22]. Moreover, attention mechanisms were investigated within Siamese-based networks [23–30]. Ref. [27] introduced an attention mechanism to capture contextual information from features at various levels, resulting in improved accuracy. SiamON [28] proposed a target-aware attention mechanism within the Siamese network, allowing the tracker to allocate more attention to the target, particularly effective in addressing occlusion challenges. The use of trackers on RGB data was explored for decades, and achieved remarkable performance [29]. Compared to HS training data, the RGB data were sufficient for training discriminative tracking networks. However, applying tracking networks designed for RGB data directly to HS object tracking tasks is not feasible due to differences in channel numbers and data distribution.

2.2. Hyperspectral Object Tracking

HS tracking has drawn growing interest recently, since the enriched spectral information gives the trackers stronger discriminative ability in demanding situations like deformation and background clutter. Based on RGB tracking methods, HS object tracking algorithms developed methods with band processing and fusion modules.

The early HS tracking approaches relied on handcrafted features and correlation filter-based trackers [1,31]. MHT [1] introduced modified HOG and global material abundance features to extract material information from HS images, and applied the correlation filter tracker BACF [32] for tracking the target. Subsequently, the authors of [3,10,33] explored different methods to enhance feature extraction from HS images, leading to further performance improvements. TSCFW introduced a spatial–spectral-weighted regularizer to inhibit the pixels dissimilar to the target and penalized unexpected peaks in the response maps [3]. Nowadays, more efforts have been devoted into deep learning methods to pursue higher performance.

Trackers based on deep learning use networks built using a transferred RGB tracking network, whereas HS training data are insufficient for training a robust tracking network [4,5]. Since RGB tracking networks operate on three-channel images, methods were investigated to reduce the channels of the HS images [13,34–36]. For instance, BAHT designed a background-aware band selection module to build images with three selected bands, which were subsequently utilized by the RGB tracking network to track the target [35]. Additionally, the bands in the HS images were regrouped into multiple three-channel false-color images [4,5,7,37,38]. SiamBAG [4] approximated group weights with a band attention module and fused the classification scores using the weights. SEE-Net studied the importance of each band in HS images with a spectral self-expressive module. The bands were regrouped into five three-channel images based on their importance. The target was then localized on the basis of the results of the five images [5]. However, these methods have notable limitations. Reducing the number of bands to three results in a significant reduction in spectral information. Furthermore, converting HS data into multiple three-channel images disrupts the spectral continuity, causing a loss of inter-band correlations that negatively impacts tracking performance. In contrast, the proposed tracker not only preserves a substantial amount of spectral information from the HS data but also captures the inter-band correlations, enabling a more robust tracking approach. This capability of retaining spectral details and learning the interactions between bands establishes our method as a more effective solution for HS tracking.

2.3. Transformer-Based Object Tracking

The Transformer model, initially introduced in NLP, was later adopted for computer vision tasks because of its remarkable performance [39–42]. For object tracking tasks, Transformer has become a prevailing method. TransT integrated features from template and search patches by leveraging Transformer-based modules, ECA and CFA [40]. Instead of the correlation operation in traditional Siamese network-based trackers, TransT exploited the Transformer-based modules' capacity for capturing long-range information, and achieved better performance. Subsequent works utilized Transformer to build the encoder and decoder, elevating the performance of Transformer-based trackers [43–47]. SiamPIN [46] introduced CNNs for extracting local information and utilized a model based on Transformer for capturing global context. Their method proposed the Trans-ConV unit block to facilitate the interaction between global and local information, resulting in significant improvement. SFTransT [47] introduced the special attention mechanism to protect high-frequency signals and achieve an all-pass filter to overcome the limitations of Transformer. Recently, researchers have developed networks that utilize Transformer blocks for feature extraction and template-search integration [48,49].

For HS object tracking, transformer was applied to fuse features of images. Implementing transformer blocks to fuse the features obtained from RGB and HS images was one of the mainstream methods [50–52]. TFTN employed 3-D convolution network to extract features from HS images, and utilized transformer-based modules to fuse features of HS images and corresponding RGB images [50]. Meanwhile, applying the transformer blocks to fuse features corresponding to different information, i.e., spectral and spatial information, was considered as another effective method [8]. Different from current transformer-based approachs in HS tracking, the approach proposed in the work modifies self-attention learning in transformer blocks to fuse the features of the false-color images regrouped from a single HS image.

3. Methods

In this section, the details of the proposed approach, MMFT, are presented. Firstly, the architecture of MMFT is illustrated. Secondly, the methodology of the proposed module, MMFF, is presented in detail. Thirdly, the proposed novel training and inference procedure of the MMFT are introduced.

3.1. Architecture of the MMFT Algorithm

The MMFT algorithm consists of four parts, i.e., band reduction, backbone, MMFF, and head network, as shown in Figure 2. Similar to the traditional Siamese network in the object tracking field, MMFT takes in a pair of images comprising the template and search patches. These patches undergo parallel processing within the network except for the head network. The head network integrates the features from the two images to predict the target's state.

3.1.1. Band Reduction

To boost the speed of the proposed tracker, multiple convolution layers are employed to reduce the channels of the input image, denoted as $X \in \mathbb{R}^{H\times W\times 16}$, to 9 channels, forming the output image $X^{'} \in \mathbb{R}^{H\times W\times 9}$. Subsequently, the output image with fewer channels is sequentially segmented into numerous false-color images, enabling the extraction of deep features by the transferred RGB tracking network. H and W represent the height and width of the HS image, respectively.

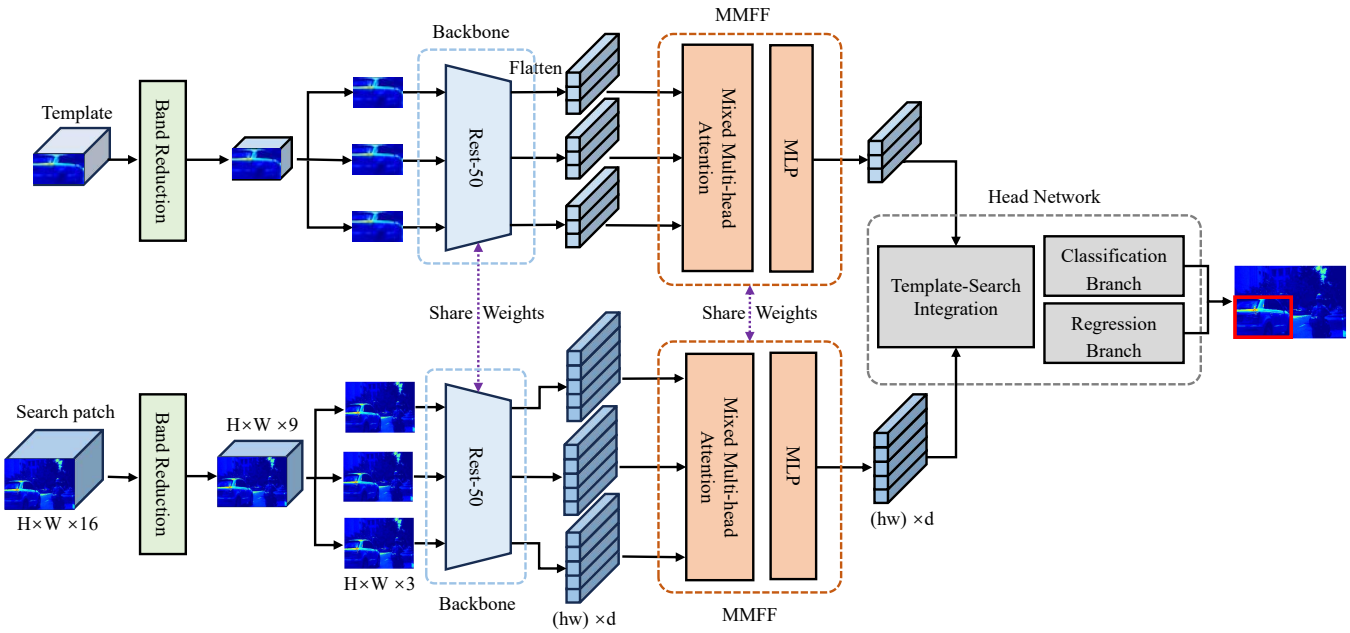

Figure 2. The pipeline of the proposed MMFT.

3.1.2. Mixed Multi-Head Feature Fusion (MMFF)

The crucial advantage of HS object tracking lies in the abundance of spectral information that enhances tracking precision in challenging scenarios. However, dividing an HS image into multiple false-color images causes the loss of interactive information between channels. To address this issue, the MMFF module is introduced to capture the interactive information among tokens within the false-color features and merge it with the features. As illustrated in Figure 2, the MMFF module is constructed for the template and search patches.

The implementation of the MMFF module is detailed in this section. As shown in Figure 3, the structure of MMFF consists of mixed multi-head attention (MMA) learning and a multilayer perceptron (MLP). Unlike the standard Transformer encoder, which uses multi-head self-attention, the MMFF module utilizes MMA to capture both intra-band dependencies within the fused feature and inter-band dependencies across different false-color features. Following the MMA block, the output fused feature is passed through an MLP to further refine its representation ability. The MLP is designed as a two-layer network: the first layer projects the feature to an intermediate dimension, using a fully connected layer followed by a ReLU activation; while the second layer projects the intermediate feature back to the original feature dimension. This design enables the model to capture more complex non-linear relationships among the fused features, thereby boosting the discriminative ability of the network. The inputs to the MMFF module are denoted as $\tilde{X}_i' \in \mathbb{R}^{hw \times d}$, and the fused feature is represented as $\tilde{X}_{fused}' \in \mathbb{R}^{hw \times d}$. The output of the module is $\tilde{X}'' \in \mathbb{R}^{hw \times d}$. Here, h and w represent the height and width of the features. d represents the channel number of the features. MMA learning serves as the key mechanism for extracting interactive information and integrating it into the fused feature.

By utilizing MMA learning, the procedures in MMFF can be formulated as follows:

$$\begin{cases} \tilde{X}'' = \sum_{i=1}^{N} w_i \tilde{X}_i' + \mathrm{Norm}(\mathrm{MMA}(\tilde{X}_1', \tilde{X}_2', \ldots, \tilde{X}_N')) \\ \tilde{X}'' = \tilde{X}'' + \mathrm{Norm}(\mathrm{MLP}(\tilde{X}'')) \end{cases}, \quad (1)$$

where MMA$(\cdot)$ represents mixed multi-head attention learning, Norm$(\cdot)$ is the normalization operator, MLP$(\cdot)$ represents the multiple layers of the perceptron, and $\tilde{X}_i^{'}$ is the ith false-color feature. Overall, MMFF fuses the false-color features and integrates the interactive information from these images into the fused feature.

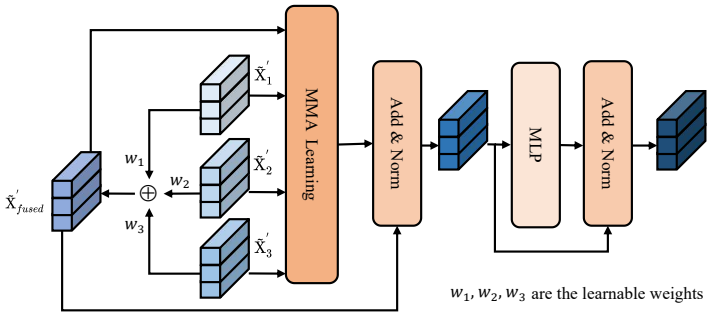

Figure 3. The structure of the mixed multi-head feature fusion module.

3.1.3. Backbone and Head Network

The backbone and head network are inspired by the RGB tracking network, which is TransT in our case. The backbone is based on a customized version of ResNet-50. In this modified version, the 5th stage has been eliminated, and the 4th stage's stride has been adjusted to 1, thereby outputting features with higher resolution. The head network consists of a template-search feature integration module and two parallel branches dedicated to classification and regression. The template-search integration module applies self-attention learning to enrich the features, and cross-attention learning to fuse the features of the template and search patches. The classification and regression branches are constructed with a three-layer perceptron. The target's location is determined by identifying the peak score in the classification map and retrieving the corresponding coordinates in the regression map.

3.2. Mixed Multi-Head Attention

MMA learning is a crucial component within the MMFF module, responsible for extracting the interactive information and integrating it into the fused feature. Further elaboration on this process is presented in the corresponding section. A detailed description of the mechanism is provided in the section. MHSA learning is commonly used in self-attention learning, since different linear projections to queries, keys, and values enhance self-attention learning [53]. MHSA is described as follows:

$$\begin{cases} \mathrm{h}_i = \mathrm{Attention}(\mathrm{Q}W_i^Q, \mathrm{K}W_i^K, \mathrm{V}W_i^V) \\ \mathrm{M} = \mathrm{Concat}(\mathrm{h}_1, \mathrm{h}_2, \ldots, \mathrm{h}_n)W^O \end{cases}, \tag{2}$$

where Attention$(\cdot)$ is the self-attention learning, Concat$(\cdot)$ represents the concatenation operation, $W_i^V \in \mathbb{R}^{d \times d_k}$, $W_i^K \in \mathbb{R}^{d \times d_k}$, $W_i^Q \in \mathbb{R}^{d \times d_k}$, and $W^O \in \mathbb{R}^{nd_k \times d}$. n is the number of heads, and $d = nd_k$ is the number of channels. Q, K, and V are the query, key, and value, which are obtained with the same feature.

In contrast to MHSA, MMA performs the self-attention and the cross-attention within one attention learning block. As illustrated in Figure 4, MMA learning consists of two groups of attention learning. One group performs self-attention on the fused feature, while the other group conducts cross-attention between the fused feature and the features of the false-color images. The self-attention learning in MMA enables the model to capture the intra-interactive information within the fused feature, while the cross-attention learning allows the model to learn the inter-interactive information between the fused feature and

the false-color features. By augmenting the learned interactive information into the fused features, MMA enhances their discriminative ability.

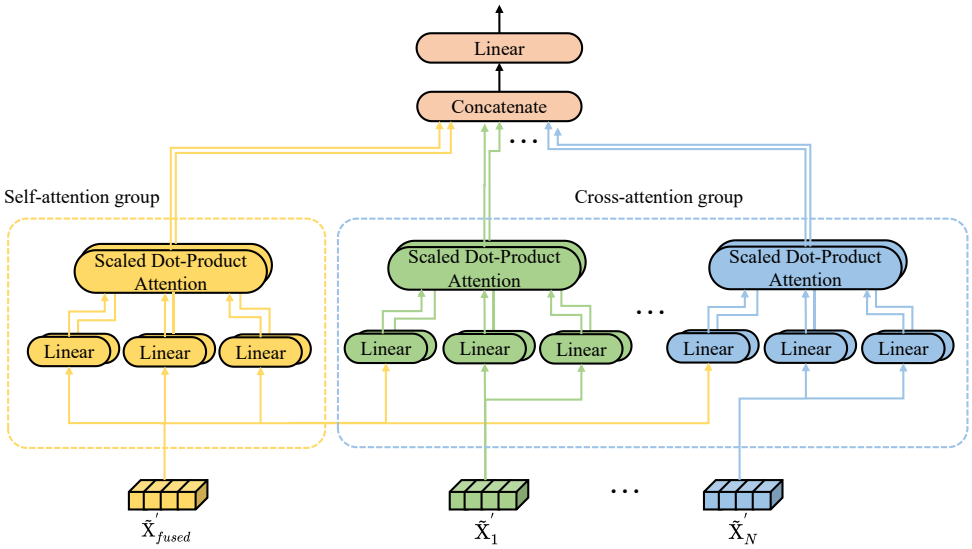

Figure 4. The implementation of MMA learning.

Specifically, the query is a fused feature calculated as follows:

$$\tilde{X}'_{fused} = \sum_{i=1}^{N} w_i \tilde{X}'_i, \tag{3}$$

$$Q = \tilde{X}'_{fused}, \tag{4}$$

where $\tilde{X}'_i$ represents the false-color features extracted from the images regrouped from one HS image, w_i is the learnable weights, and N is the number of false-color images. Then, for different h_i, the keys and values are a group of features, which consists of the fused feature and the false-color features as follows:

$$\begin{cases} K^F = \tilde{X}'_{fused} \\ K_i^X = \tilde{X}'_i \\ V^F = \tilde{X}'_{fused} \\ V_i^X = \tilde{X}'_i \end{cases}, \tag{5}$$

where K^F and V^F represent the key and value corresponding to the fused feature, and K_i^X and V_i^X denote the keys and values corresponding to the false-color features. The attention learning can be described as follows.

$$h_i = \text{softmax}\left(\frac{(q_i + p)(k_i + p)^T}{\sqrt{d_k}}\right) v_i, \tag{6}$$

where $q_i = QW_i^Q$, $k_i = KW_i^K$, and $v_i = VW_i^V$. K consists of K^F and K_i^X, and V has the same situation. d_k is the number of channels in k_i. p represents the positional embedding information, which includes the positional relationships between locations within the feature map. In Formula (6), the position embeddings for the queries and keys are the same because the positional embedding corresponds to the size of the features [42], and the fused feature and the false-color features, which generate the queries and keys, share the same

size. In this way, MMFF realizes the fusion of false-color features and augments interactive information of tokens into the fused feature in one MMA learning procedure.

3.3. Two-Step Training Procedure

The MMFT network consists of two types of modules, those transferred from the RGB tracking network and those specifically built for HS object tracking. Modules originating from the RGB tracking network have already been trained on large datasets of RGB tracking, acquiring discriminative ability for tracking RGB objects. Since tracking the object with the false-color images shares most of the prior knowledge on feature extraction and movement state estimation, the components transferred from the RGB tracking network are frozen during the HS training process. Modules constructed for HS object tracking, such as the band reduction and MMFF modules, derive their parameters from the HS data. However, the HS data are insufficient for training the parameters in the modules, especially for the Transformer-based MMFF. To obtain the parameters with enhanced generalization and adaptability for the HS object tracking task, a two-step training procedure is proposed to train the modules, as shown in Figure 5. In step 1, modules designed for HS object tracking, i.e., the band reduction module and the MMFF module, are pre-trained on a large RGB dataset to improve generalization. In step 2, the same modules in the MMFT network are fine-tuned on the limited HS data to adapt to the HS tracking task. During both steps of training, modules transferred from the RGB tracking network are frozen. The implementation details of the TS training procedure are outlined in Algorithms 1 and 2, which provide a detailed description of the first and second steps of the training process. The large RGB dataset in step 1 is the commonly used object tracking dataset, GOT-10K, which contains more than 10 K sequences, while the HS training dataset in step 2 consists of only 40 sequences. Therefore, the approach increases the amount of training data.

Algorithm 1: The first step of the TS training procedure

Input: N frames of RGB datasets $X_{RGB} \in \mathbb{R}^{H\times W\times 3}$, where W, H, and C represent the width, height, and channel of the RGB image, respectively.

Output: The pre-trained weight of the parameters in the MMFT.

Initialization: Load the weights from the RGB tracker into the MMFT, initialize the parameters in the Band Reduction and MMFF modules with random initialization, and freeze the parameters in the Backbone and Head network;

for $n = 1, 2, \ldots, N$ **do**

- Concatenate X_{RGB} along the channel dimension to form $X'_{RGB} \in \mathbb{R}^{H\times W\times 3N}$;
- Put X'_{RGB} into the Band Reduction to obtain $X' \in \mathbb{R}^{H\times W\times 9}$;
- Divide X' into multiple false-color images $X'_1, X'_2, X'_3 \in \mathbb{R}^{H\times W\times 3}$;
- Put X'_1, X'_2, X'_3 into the Backbone to obtain $\tilde{X}'_1, \tilde{X}'_2, \tilde{X}'_3 \in \mathbb{R}^{hw\times d}$;
- Put $\tilde{X}'_1, \tilde{X}'_2, \tilde{X}'_3$ into the MMFF to obtain $\tilde{X}'' \in \mathbb{R}^{hw\times d}$;
- Put $\tilde{X}''$ into the Head network to obtain the classification and regression responses ;
- Calculate the L_{cls} using the p_j in the classification response and the y_j in the groundtruth;
- Calculate the L_{reg} using the b^j in the regression response and the $\hat{b^j}$ in the groundtruth ;

end for

Algorithm 2: The second step of the TS training procedure

Input: N frames of the HS dataset $X_{HS} \in \mathbb{R}^{H\times W\times 16}$.

Output: The fine-tuned weight of the parameters in the MMFT.

Initialization: Load the pre-trained weights from the first step into the MMFT and freeze the parameters in the Backbone and Head network;

for $n = 1, 2, \ldots, N$ **do**

  Put X_{HS} into the Band Reduction to obtain $X^{'} \in \mathbb{R}^{H\times W\times 9}$;

  Divide $X^{'}$ into multiple false-color images $X_1^{'}, X_2^{'}, X_3^{'} \in \mathbb{R}^{H\times W\times 3}$;

  Put $X_1^{'}, X_2^{'}, X_3^{'}$ into the Backbone to obtain $\tilde{X}_1^{'}, \tilde{X}_2^{'}, \tilde{X}_3^{'} \in \mathbb{R}^{hw\times d}$;

  Put $\tilde{X}_1^{'}, \tilde{X}_2^{'}, \tilde{X}_3^{'}$ into the MMFF to obtain $\tilde{X}^{''} \in \mathbb{R}^{hw\times d}$;

  Put $\tilde{X}^{''}$ into the Head network to obtain the classification and regression responses ;

  Calculate the L_{cls} using the p_j in the classification response and the y_j in the groundtruth;

  Calculate the L_{reg} using the b^j in the regression response and the $\hat{b^j}$ in the groundtruth ;

end for

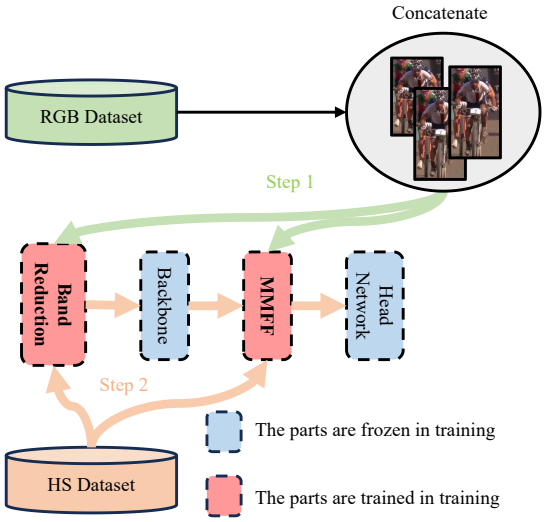

Figure 5. Illustration of the two-step training procedure.

The existing two-step training strategy in L2RCF [54], which trains the tracking network with labeled remote sensing (RS) data in the first step and RS data with generated pseudo-labels in the second step, differs from the TS training procedure in MMFT in three key aspects. Firstly, MMFT uses cross-modal data for training, while L2RCF utilizes intra-domain data. Secondly, in MMFT, only the HS-specific modules are trained, whereas L2RCF trains the full network parameters. Thirdly, in step 2, MMFT focuses on fine-tuning the modules with HS data to adapt the model for the HS tracking task, while L2RCF generates pseudo-labels to augment the training set and enhance the generalization of the classifier.

Prompt learning is the commonly used method for handling data scarcity by adding small model branches or adapters, and fine-tuning only the added parameters. Different from prompt learning, the TS training procedure offers a more effective solution. The TS training procedure addresses the scarcity of training data by pre-training the additional modules for the HS object tracking task on the abundant RGB dataset in step 1. Then, the model is fine-tuned on HS data in step 2 to adapt the model specifically for the HS tracking task. Moreover, prompt learning cannot be directly applied to the HS object tracking task.

The method does not change the structure of the base model, which in this work is the RGB object tracking network. To adapt it for the HS object tracking task, the HS images need to be transferred into three-channel images to match the base model. This operation differs from the proposed TS training strategy and leads to the loss of the spectral information.

The RGB data contain three channels, which is not suitable for the MMFT network. Hence, the RGB images, $X_{RGB} \in \mathbb{R}^{H\times W\times 3}$, are concatenated along the channel dimension to form the images, $X'_{RGB} \in \mathbb{R}^{H\times W\times 3N}$, making them suitable for training the MMFT algorithm. The RGB images are sampled from the large RGB tracking dataset. After training with the RGB data, the MMFT is fine-tuned on the HS dataset. The image pairs, the template and the search patches, are sampled from one sequence to collect the training samples. And the same loss functions are utilized for the TS training. The classification branch employs the binary cross-entropy loss function, which is calculated as follows:

$$L_{cls} = -\sum_{j}[y_j \log(p_j) + (1-y_j)\log(1-p_j)], \tag{7}$$

where y_j represents the foreground or background with 1 and 0, and p_j is the probability of the sample belonging to the foreground. The generalized IoU (GIoU) loss and L_1 loss [55] are applied in the bounding box regression branch, and can be written as follows:

$$L_{reg} = \sum_{j} 1_{\mathrm{IoU}(b^j,\hat{b^j})>0}[p_1 L_{\mathrm{GIoU}}(b^j,\hat{b^j}) + p_2 L_1(b^j,\hat{b^j})], \tag{8}$$

where the two hyperparameters p_1 and p_2 control the relative importance of the GIoU loss and the L_1 loss [40]. p_1 is critical for handling large misalignments between the prediction and groundtruth, while p_2 is effective for fine-tuning the bounding box coordinates once the prediction is close to the groundtruth bounding box. In this work, p_1 and p_2 are set to 2 and 5, respectively. b^j denotes the predicted bounding box, while $\hat{b^j}$ corresponds to the regression target.

3.4. Inference

In the inference procedure of MMFT, the initial frame of a video is applied to obtain the template patch based on the given location and the state of the target, and the template patch is used in the following frames of the video. The search patch in the following frames is obtained according to the model's predictions in the previous frame. The MMFT network produces two score maps: the classification map and the regression map. The classification map contains two channels, indicating the likelihood of the sample belonging to the foreground or background. And four channels in the regression map represent the normalized coordinates. To account for the proximity of target locations in consecutive frames, a weighted mask is used to penalize scores that deviate significantly from the center within the classification map.

4. Experiments

The section commences with an overview of the experimental setup, progressing to an ablation study aimed at validating the impacts of the proposed components on the MMFT and TS training process. Subsequently, quantitative and qualitative analyses are presented to compare MMFT with other RGB and HS trackers.

4.1. Experimental Setup

4.1.1. Dataset

Due to the TS training procedure, which first pre-trains the designed modules on the RGB dataset to gain better generalization, and then fine-tunes them on the HS dataset

to adapt to the tracking task in HS videos, both RGB and HS datasets are utilized. The GOT-10K dataset is chosen as the RGB dataset and the HS dataset is provided in [1]. The HS dataset contains three types of videos, i.e., HS videos, false-color videos, and RGB videos. The false-color videos are obtained from the HS videos with the CIE color matching function [1], which provides a standardized method to convert spectral power distributions into color coordinates that approximate human color perception. The RGB videos are shot at the same time and at a similar angle to the HS videos. Hence, the three types of videos describe almost the same scenarios, which is suitable for comparing different trackers. There are 35 videos for testing and 40 videos for training. These videos are classified based on 11 challenging factors, i.e., out-of-plane rotation (OPR), out of view (OV), background clutter (BC), in-plane rotation (IPR), fast motion (FM), motion blur (MB), deformation (DEF), occlusion (OCC), scale variation (SV), illumination variation (IV), and low resolution (LR). The trackers' performance is assessed through precision plots, success plots, as well as metrics such as the area under the curve (AUC) score and distance precision (DP) score. The AUC score is calculated from the success plot data. The DP score is obtained by counting the reported results at a 20-pixel threshold.

4.1.2. Implementation Details

In the experiments, MMFT is coded in Python using PyTorch. The training and evaluations are conducted on a computer with a Xeon Silver 4210R CPU and two NVIDIA RTX 3090 GPUs. The RGB tracking network utilized for transferring is TransT, which is trained on large RGB datasets, including, COCO [56], LaSOT [57], GOT-10K [58], and TrackingNet [59]. During the TS training procedure, all module parameters, except those of the band reduction module and MMFF, are kept frozen. In this work, the parameters are updated using the AdamW optimizer. For the first step of training, a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4} are applied. The training process is conducted for 100 epochs, with a batch size of 32. The learning rate is scheduled to decrease by a factor of 10 after 60 epochs. For the second step of training, a learning rate of 1×10^{-5} and a weight decay of 1×10^{-4} are applied. The training process is conducted for 30 epochs, with a batch size of 32. The learning rate is scheduled to decrease by a factor of 10 after 20 epochs. Moreover, the MMFF architecture is configured with 8 heads for the experiments. The sizes of the inputs, i.e., template and search patches, are resized to 128×128 and 256×256, respectively.

5. Results and Analysis

5.1. Ablation Study

5.1.1. Effectiveness of the Band Reduction

To validate the effectiveness of the band reduction, an ablation study on the band reduction with different numbers of output channels is conducted. As shown in Table 1, the tracker with an output setting of 15 channels achieves the best tracking performance, with AUC and DP@20P scores of 0.679 and 0.917, respectively. However, its inference speed is limited to 21.1 FPS. When the number of output channels is reduced to nine, the tracker achieves AUC and DP@20P scores of 0.675 and 0.914, respectively, while improving the inference speed to 27.9 FPS. Reducing the number of channels to six achieves the worst performance in Table 1 due to the loss of spectral information. Considering the trade-off between inference speed and performance, the number of output channels is set to nine for the band reduction.

Table 1. Comparison of band reduction with different numbers of output channels.

Output Channels				AUC	DP@20P	FPS
6	9	12	15			
✓				0.586	0.826	**30.9**
	✓			0.675	0.914	**27.9**
		✓		**0.676**	**0.915**	24.5
			✓	**0.679**	**0.917**	21.1

The top two values are marked in red and blue.

5.1.2. Superiority of Proposed MMFF

To highlight the superiority of the MMFF module, a comprehensive experiment is conducted, and the results are presented in Table 2. In the experiment, the number of channels of the input images for all methods are reduced from 16 to 9 using the band reduction module. In Table 2, the baseline tracker utilizes simple addition to fuse the features, while the weighted fusion tracker employs learnable weights to fuse the features. The concatenated tracker concatenates the features and reduces them to the original channel dimensions using a CNN. The self-attention weighted fusion tracker calculates self-attention on the fused feature, which is obtained by applying weighted fusion to combine false-color features. Concatenated self-attention fusion involves merging the enhanced concatenated feature using learnable weights. An enhanced concatenated feature is obtained by applying self-attention to the concatenated feature, which is obtained by merging features along the channel dimension for the false-color features. Additionally, MMFF_na and MMFF are trackers that perform MMA computations using the method illustrated in Figure 4. They subsequently fuse the features according to Equation (1). During the MMA process, all queries are generated from the fused feature. Specifically, MMFF_na generates sets of keys and values in a 5:1:1:1 ratio from the fused feature and the false-color features. MMFF generates sets of keys and values in a 2:2:2:2 ratio for the same features.

Table 2. Quantitative analysis of different feature fusion modules.

Model	AUC	DP@20P	Params (M)	FLOPs (G)
Baseline	0.586	0.820	18.54	43.48
Concatenated	0.548	0.801	21.69	47.51
Weighted fusion	0.675	0.914	18.54	43.48
Self-attention weighted fusion	0.677	0.916	31.15	59.62
Concatenated self-attention fusion	0.682	0.915	62.61	99.89
MMFF_na	**0.685**	**0.917**	37.44	67.67
MMFF	**0.689**	**0.919**	37.44	67.67

The top two values are marked in red and blue.

As shown in Table 2, the baseline tracker achieves an AUC score of 0.586 and a DP@20P score of 0.820. The performance of the concatenated tracker is inferior to the baseline due to information loss incurred during dimension reduction. Compared with the baseline, the weighted fusion method yields improvement in the tracking results, enhancing the tracker's performance by 0.089 and 0.094 in terms of AUC and DP@20P scores, respectively. The performance enhancement can be attributed to the contribution of the features corresponding to false-color images, which enhances the overall tracking performance. The self-attention weighted fusion calculates attention on the weighted feature, further enhancing it and resulting in better tracking performance. The concatenated self-attention fusion yields a superior AUC score due to the enhanced interaction between false-color features through the application of self-attention to the concatenated feature. However, this improvement comes at the expense of increased computational resources. Concatenating features triples the number of feature tokens, resulting in a nine-fold increase

in computational cost for calculating self-attention. Therefore, the method is inefficient. As depicted in Table 2, the proposed MMFF and MMFF_na modules rank first and second in the experiment. Replacing the feature fusion module with MMFF and MMFF_na, the tracking performance significantly improves compared to the baseline, achieving enhancements of 0.103 and 0.099 in AUC score, respectively. These results demonstrate the effectiveness of the proposed module.

Table 2 also presents the model complexity and computational cost of the various methods. The top three methods in Table 2 exhibit similar model complexity and computational cost, as they employ simple data processing techniques, i.e., element-wise addition and convolution layers. In contrast, the model complexity and computational cost of the methods in the last four lines of Table 2 are significantly higher than those in the top three lines due to application of the attention mechanism. Compared to weighted fusion, self-attention weighted fusion contributes to a limited improvement in the performance, with increased model complexity and computational cost. Concatenated self-attention fusion has the largest number of parameters and FLOPs since the self-attention is conducted on the large concatenated feature. The size of the concatenated feature is three times that of the features in MMFF_na and MMFF. MMFF_na and MMFF have the same number of parameters and FLOPs since they both apply MMA to compute the attention. Overall, MMFF achieves the best performance with a limited increase in the model complexity and computational cost.

5.1.3. Impact of Different Training Strategies

An experiment is conducted to confirm the superiority of the proposed training procedure. As there is a significant shortage of hyperspectral video data, an RGB dataset is utilized for pre-training the band reduction and MMFF modules, and an HS dataset is applied for fine-tuning the modules. As described in Section 3.3, since the MMFT network requires input images with more than three channels, the original RGB images, $X_{RGB} \in \mathbb{R}^{H\times W\times 3}$, are modified by concatenating N copies along the channel dimension to form images, $X'_{RGB} \in \mathbb{R}^{H\times W\times 3N}$. This straightforward transformation ensures that the RGB data can be effectively utilized during the pre-training stage, compensating for the channel mismatch between the RGB and hyperspectral data. Table 3 presents the experimental results comparing different training procedures. RGB training refers to the training procedure involving training the modules using the modified RGB data only. HS training indicates the training procedure where the modules are trained exclusively with the HS training dataset. TS training is the proposed training procedure, involving the pre-training of modules using RGB data followed by fine-tuning with HS data. As illustrated in Table 3, conducting RGB training on the modules using the GOT-10K dataset results in a competitive performance, with AUC and DP@20P scores reaching 0.679 and 0.910, respectively. However, when performing RGB training with the LaSOT dataset or the combination of LaSOT and GOT-10K, the performances of the trackers drop. Based on the result of RGB training on the GOT-10K dataset, conducting TS training on the proposed modules leads to the best performance, achieving AUC and DP@20P scores of 0.689 and 0.919, respectively. Conducting HS training of these modules results in a decrease in tracking performance, with a 0.016 drop in the AUC score compared to the RGB training on the GOT-10K dataset. This decline is attributed to the limited availability of HS data, which leads to overfitting during training. In summary, the TS training procedure for the modules that cannot be transferred from the existing RGB tracking network proves its effectiveness in HS tracking, achieving superior performance compared to other training procedures.

Table 3. Quantitative analysis of the different training strategies.

Training Dataset			Training Strategy			AUC	DP@20P
GOT-10K	LaSOT	HSI	RGB Training	HS Training	TS Training		
✓			✓			0.679	0.910
	✓		✓			0.674	0.907
✓	✓		✓			0.677	0.903
		✓		✓		0.665	0.904
✓		✓			✓	0.689	0.919

The top two values are marked in red and blue.

To demonstrate the occurrence of overfitting during HS training, Figure 6 presents the training loss and performance curves. As shown in Figure 6, the training loss consistently decreases over the epochs, indicating that the model is fitting the training data. However, the AUC performance on the testing dataset peaks early and then declines. This performance drop after the initial improvement confirms that the model is overfitting to the training data, as it fails to generalize well to unseen data.

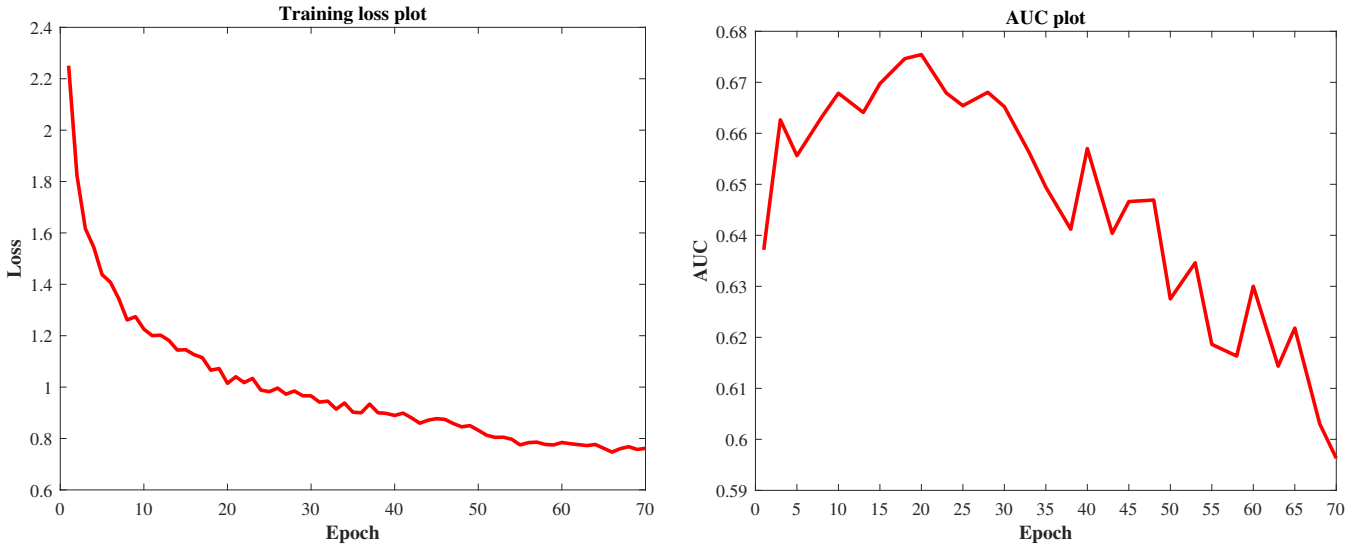

Figure 6. Plots of training loss and performance over the epochs.

5.2. Comparison with State-of-the-Art RGB Trackers

To evaluate the performance of the proposed tracker, a competitive comparison with eleven RGB trackers is conducted, including fDSST [60], SRDCF [61], BACF [32], MCCT [62], SiamCAR [19], SiamRPN++ [18], TransT [40], SwinTrack [48], SeqTrack [63], ARTrack [64], and HIPTrack [65]. Notably, HIPTrack and ARTrack have achieved SOTA performance on RGB tracking datasets. The performance of all the compared trackers is displayed in Figures 7 and 8. MMFT performs inference on HS videos. The other trackers in Figure 7 perform inference on false-color videos, and those in Figure 8 perform inference with RGB videos.

As demonstrated in Figures 7 and 8, MMFT outperforms other RGB trackers in precision and success, achieving superior performance at any threshold. It attains an AUC score of 0.689 and a DP@20P score of 0.919. Moreover, a comparison between Figures 7 and 8 reveals that RGB trackers exhibit better tracking results on RGB data in contrast to false-color data. Notably, ARTrack, a state-of-the-art tracker that excels on RGB object tracking datasets and ranks among the top three trackers for both RGB and false-color videos on the HS dataset, experiences a significant performance drop when applied to false-color videos, with decreases of 0.024 in AUC and 0.018 in DP@20P scores. The observations emphasize

the limitations of RGB-based trackers when applied to HS videos, particularly when HS videos are transformed into three-channel formats. To address this, MMFT is specifically designed for tracking in HS videos. The results presented in Figure 7 demonstrate that MMFT outperforms ARTrack by a significant margin, surpassing it by more than 0.068 in AUC and 0.023 in DP@20P scores, further validating its effectiveness for HS tracking. This enhancement can be attributed to the effective utilization of spectral information by MMFT in HS videos. The abundance of spectral information in HS videos offers more discriminative cues, leading to improved tracking performance and contributing to the robustness.

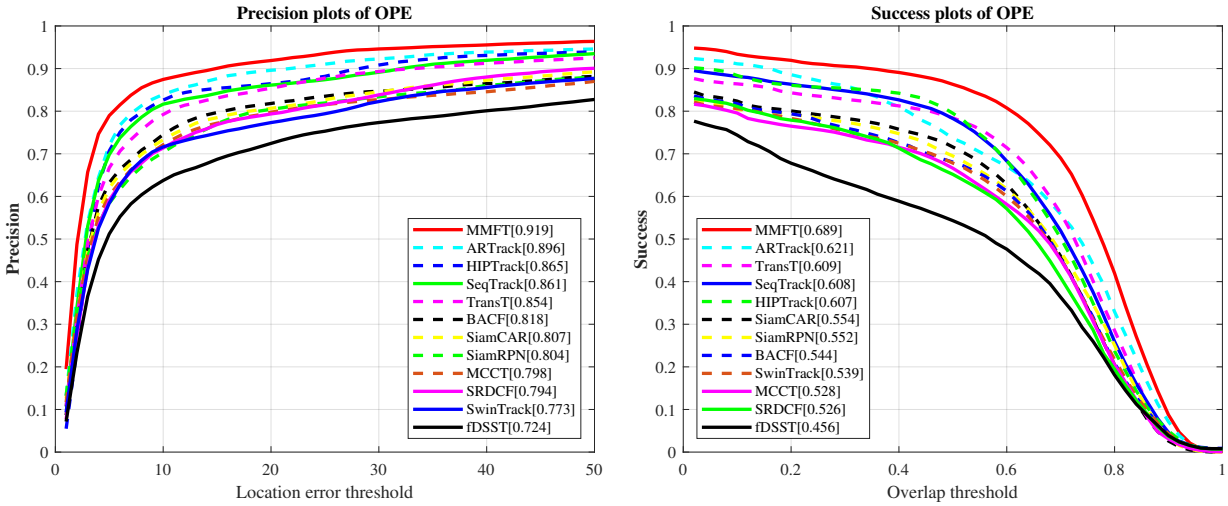

Figure 7. Comparisons of MMFT and trackers on the corresponding false-color videos.

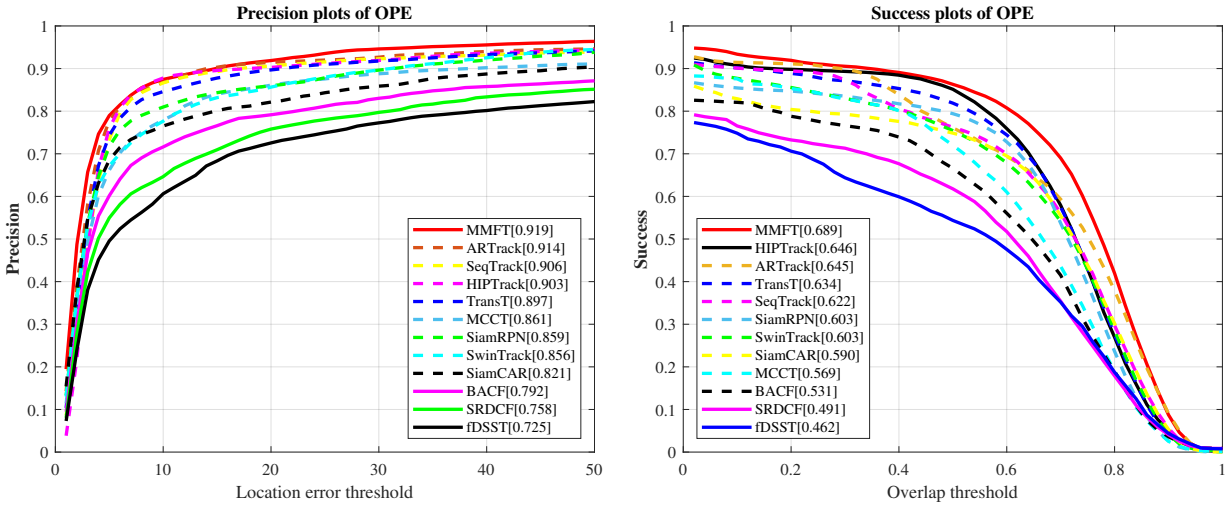

Figure 8. Comparisons of MMFT and trackers on the corresponding RGB videos.

5.3. Comparison with Hyperspectral Trackers

To reveal the superior tracking performance of MMFT, a comparative experiment is conducted against twelve SOTA HS trackers, i.e., DeepHKCF [34], MFI-HVT [66], MHT [1], SST-Net [14], BAE-Net [37], SSDT-Net [13], SiamHYPER [2], SiamHT [67], SEE-Net [5], TBR-Net [68], PHTrack [69], and SPIRIT [70]. The results, summarized in Table 4, clearly show that MMFT outperforms the other trackers, achieving the highest AUC score of 0.689. Among the comparative trackers, DeepHKCF, an adaptation of the KCF tracker, fails to incorporate spectral–spatial structural information from HS data, leading to significantly degraded performance. MHT relies on handcrafted features derived from HS data for tracking. However, these features do not adequately capture the rich information present in

HS videos, resulting in a lower AUC score compared to MMFT. Methods such as MFI-HVT, SiamHT, and SSDT-Net convert HS videos into three-channel false-color videos before extracting features for tracking. These methods result in a substantial loss of hyperspectral spectral information, which adversely affects tracking accuracy and results in inferior performance compared to MMFT. SST-Net, BAE-Net, SEE-Net, and TBR-Net rearrange HS videos based on the importance of spectral bands, producing multiple false-color videos as inputs to the tracker. Similarly, PHTrack generates such videos using a neural network. Although these methods retain a significant portion of spectral information in HS videos, the lack of feature interaction within the false-color videos impedes tracking performance, leading to a decrease in the AUC score compared to the proposed method. SiamHYPER, which utilizes both RGB and HS videos, achieves an AUC score of 0.678. The HS data in this approach serve as supplementary information to assist in RGB video tracking. SPIRIT, which employs a template update mechanism, achieves an AUC score of 0.679. However, the performance of MMFT still surpasses it, with an AUC score 0.010 higher. MMFT's superior performance is attributed to several critical factors. First, the MMFF module facilitates effective feature interaction across false-color videos, resulting in improved tracking accuracy. Furthermore, pre-training MMFT on a large dataset of RGB videos enhances its generalization capability, contributing to its overall superior performance.

Table 4 further demonstrates the inference speeds of MMFT and the compared HS trackers. MMFT achieves an inference speed of 26.1 FPS, enabling real-time tracking. Among the trackers evaluated, MMFT is the second fastest, only surpassed by SSDT-Net. However, SSDT-Net compromises spectral information through the reduction of bands in the HS data, resulting in a significant decrease in tracking precision. Considering tracking performance and inference speed, MMFT offers a better balance of accuracy and practicality.

Table 4. Performance comparison with hyperspectral trackers of AUC and FPS.

Tracker	**DeepHKCF [34]**	**MHT [1]**	**MFI-HVT [66]**	**BAE-Net [37]**	**SST-Net [14]**	**SSDT-Net [13]**	**SiamHYPER [2]**
AUC	0.328	0.588	0.604	0.606	0.623	0.639	0.678
FPS	0.91	2.61	2.42	0.72	0.65	**35.7**	19.0
Tracker	**SiamHT [67]**	**SEE-Net [5]**	**TBR-Net [68]**	**PHTrack [69]**	**SPIRIT [70]**	**MMFT**	
AUC	0.621	0.666	0.660	0.660	**0.679**	**0.689**	
FPS	16.0	8.72	14.9	15.2	26.0	**26.1**	

The top two values are marked in red and blue.

5.4. Attribute-Based Evaluation

This section uses 11 attributes to evaluate MMFT and seven compared trackers, i.e., SiamCAR [19], SiamRPN++ [18], and SwinTrack [48], and four HS trackers, i.e., MHT [1], SiamHYPER [2], BAE-Net [37], and SSDT-Net [13]. Table 5 shows the trackers' performance in AUC score. The results presented in Table 5 reveal that MMFT achieves a top-two position across 10 out of the 11 attributes, with the exception of OV. Notably, in attributes such as BC and OCC, the performance of HS trackers outperforms RGB trackers due to the enriched spectral information present in HS videos. Notably, for the IPR and MB attributes, the MMFT demonstrates superiority over the second-ranked SiamHYPER tracker, with an AUC score gain of 0.028 and 0.020, respectively. Due to the MMFT effectively utilizing valuable feature information in HS videos, it provides robust performance in addressing challenges related to in-plane rotation and motion blur.

Table 5. Attribute-based comparison on hyperspectral videos or the corresponding false-color videos.

Attribute	MMFT	SiamCAR	SiamRPN++	SSDT-Net	SiamHYPER	BAE-Net	MHT	SwinTrack
Background clutter (BC)	**0.710**	0.530	0.587	0.663	**0.702**	0.631	0.594	0.456
Deformation (DEF)	**0.735**	0.719	0.684	0.685	**0.720**	0.679	0.664	0.707
Fast motion (FM)	**0.695**	0.665	0.559	0.603	**0.710**	0.607	0.541	0.493
In-plane rotation (IPR)	**0.748**	0.631	0.674	0.666	**0.720**	0.699	0.670	0.567
Illumination variation (IV)	**0.587**	0.417	0.478	0.543	**0.592**	0.440	0.474	0.489
Low resolution (LR)	**0.686**	0.435	0.475	0.521	**0.664**	0.491	0.478	0.490
Motion blur (MB)	**0.770**	0.627	0.539	0.579	**0.750**	0.594	0.560	0.634
Occlusion (OCC)	**0.630**	0.527	0.544	0.607	**0.635**	0.555	0.565	0.533
Out-of-plane Rotation (OPR)	**0.737**	0.654	0.697	0.695	**0.706**	0.693	0.631	0.619
Out of view (OV)	0.685	0.602	0.608	**0.732**	0.596	0.516	0.620	**0.710**
Scale variation (SV)	**0.668**	0.548	0.591	0.639	**0.657**	0.608	0.564	0.586

The top two values are marked in red and blue.

5.5. Qualitative Analysis of Visual Tracking Results

To comprehensively verify the advantages of MMFT, a qualitative evaluation of various trackers is provided. The evaluation includes RGB trackers like SiamCAR [19], SwinTrack [48], and TransT [40], and HS trackers such as SSDT-Net [13] and BAE-Net [37]. The visualization results of all the trackers in Figure 9 illustrate the performance on the Basketball, Coke, Fruit, Paper, and Rider2 image sequences from top to bottom. These visualizations clearly indicate that the HS tracker demonstrates notably superior performance compared to the RGB tracker. This can be attributed to the capacity of HS trackers to utilize spectral information, enabling them to address challenges like background clutter and occlusion. SSDT-Net suffers from a loss of spectral information due to its band fusion method, resulting in poor tracking performance when faced with low-resolution and occlusion scenarios, as in the cases of Basketball and Rider2. BAE-Net effectively utilizes the spectral information from the HS video. However, its inability to fuse features across false-color videos results in target loss in sequences like Basketball and Fruit. In contrast to the compared trackers, MMFT accurately tracks the target in all five sequences. This achievement is attributed to the efficient processing of hyperspectral information by the MMFF module. Furthermore, the results emphasize the capability of MMFT to handle challenging scenarios, including background clutter (Coke, Fruit, Paper), low resolution (Basketball, Rider2), occlusion (Basketball, Fruit, Rider2), and fast motion (Basketball, Coke).

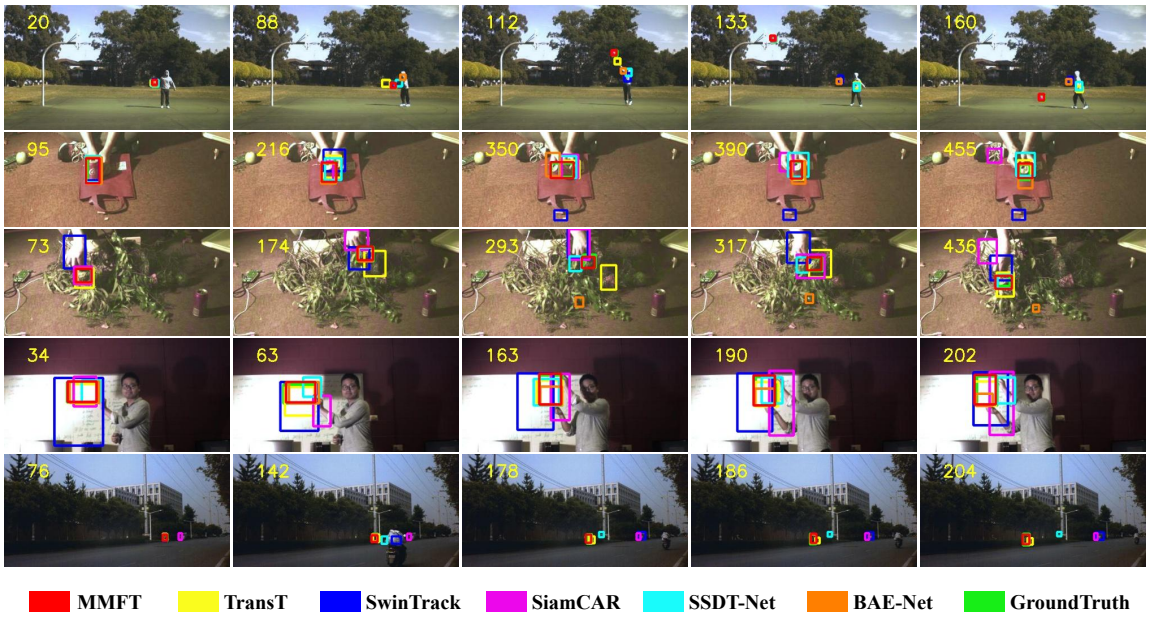

Figure 9. Visualization results of proposed MMFT tracker compared with several trackers.

6. Discussion

While this work demonstrates that MMFT achieves state-of-the-art performance in HS object tracking, several limitations merit discussion and will inform our future research directions.

Firstly, the decision to reduce the HS images from 16 to 9 channels represents a compromise between computational efficiency and tracking performance. Although the experimental results in Table 1 indicate that this reduction achieves a favorable balance, it may lead to the loss of critical spectral details in certain scenarios, thereby affecting feature extraction and overall tracking performance. To address this, we plan to investigate adaptive spectral band selection methods that can dynamically determine the optimal number of channels for data with abundant spectral information.

Secondly, the results in Table 3 demonstrate that the TS training procedure is effective for the HS tracking task. However, this approach is sensitive to domain shifts. Specifically, if the modified RGB data used in the first step do not adequately capture the spectral features of the HS domain, the subsequent fine-tuning step may not fully bridge the domain gap, resulting in limited performance improvements. Therefore, training the network solely on HS data would be preferable. However, the large number of parameters in the HS tracking network can lead to overfitting when only limited HS data are available. There are two potential solutions to address the issue: increasing the number of HS tracking data and reducing the number of trainable parameters in the tracking network. The former method fundamentally addresses the issue, such as using modified RGB data as pseudo-HS data in the first step of the TS training procedure for model training. The latter method indirectly handles overfitting via parameter-efficient fine-tuning techniques that introduce only a small number of additional parameters into the pre-trained model. This technique, which allows effective adaptation with limited HS data, will be a primary focus of our future research efforts.

7. Conclusions

A new hyperspectral object tracking method, MMFT, is proposed, which incorporates a novel feature-level fusion using mixed multi-head attention within the tracking network. To enhance the feature representation capacity, a feature fusion module based on MMA is proposed. This module, MMFF, integrates features from false-color images regrouped from a single HS image and incorporates the learned interactive information between different false-color images into the fused feature. Furthermore, to address the challenges posed by limited HS training data, the TS training procedure for hyperspectral object tracking network is introduced. The procedure involves pre-training the modules designed for hyperspectral object tracking using an extensive set of modified RGB data to improve generalization, followed by fine-tuning on a limited dataset of HS data for the task adaption. The experimental results demonstrate the effectiveness of the MMFF and TS training approach in improving tracking performance. MMFT achieves an AUC score of 0.689, with an inference speed of 26.1 FPS.

Author Contributions: Conceptualization, L.G., L.C., and W.X.; methodology, L.G. and L.C.; software, L.C.; validation, L.C.; formal analysis, L.G. and L.C.; data curation, L.C.; writing—original draft preparation, L.G. and L.C.; writing—review and editing, L.G., L.C., Y.J., and B.X.; funding acquisition, L.G., W.X., and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the project CEIEC-2022-ZM02-0247 and the Fundamental Research Funds for the Central Universities under Grant ZYTS25269.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: In this work, the dataset utilized in the experiments is obtained from https://www.hsitracking.com/ (accessed on 5 April 2021).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xiong, F.; Zhou, J.; Qian, Y. Material based object tracking in hyperspectral videos. *IEEE Trans. Image Process.* **2020**, *29*, 3719–3733. [CrossRef] [PubMed]
2. Liu, Z.; Wang, X.; Zhong, Y.; Shu, M.; Sun, C. SiamHYPER: Learning a Hyperspectral Object Tracker From an RGB-Based Tracker. *IEEE Trans. Image Process.* **2022**, *31*, 7116–7129. [CrossRef] [PubMed]
3. Hou, Z.; Li, W.; Zhou, J.; Tao, R. Spatial–Spectral Weighted and Regularized Tensor Sparse Correlation Filter for Object Tracking in Hyperspectral Videos. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–12. [CrossRef]
4. Li, W.; Hou, Z.; Zhou, J.; Tao, R. SiamBAG: Band Attention Grouping-based Siamese Object Tracking Network for Hyperspectral Videos. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–12. [CrossRef]
5. Li, Z.; Xiong, F.; Zhou, J.; Lu, J.; Qian, Y. Learning a Deep Ensemble Network with Band Importance for Hyperspectral Object Tracking. *IEEE Trans. Image Process.* **2023**, *32*, 2901–2914. [CrossRef]
6. Ouyang, E.; Wu, J.; Li, B.; Zhao, L.; Hu, W. Band Regrouping and Response-Level Fusion for End-to-End Hyperspectral Object Tracking. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
7. Gao, L.; Liu, P.; Jiang, Y.; Xie, W.; Lei, J.; Li, Y.; Du, Q. CBFF-Net: A New Framework for Efficient and Accurate Hyperspectral Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
8. Wang, Y.; Liu, Y.; Ma, M.; Mei, S. A Spectral–Spatial Transformer Fusion Method for Hyperspectral Video Tracking. *Remote Sens.* **2023**, *15*, 1735. [CrossRef]
9. Qian, K.; Zhou, J.; Xiong, F.; Zhou, H.; Du, J. Object tracking in hyperspectral videos with convolutional features and kernelized correlation filter. In *Smart Multimedia, ICSM 2018*; Springer: Cham, Switzerland, 2018; pp. 308–319.
10. Tang, Y.; Liu, Y.; Huang, H. Target-aware and spatial-spectral discriminant feature joint correlation filters for hyperspectral video object tracking. *Comput. Vis. Image. Underst.* **2022**, *223*, 103535. [CrossRef]
11. Wang, S.; Qian, K.; Chen, P. BS-SiamRPN: Hyperspectral Video Tracking based on Band Selection and the Siamese Region Proposal Network. In Proceedings of the 2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Rome, Italy, 13–16 September 2022; pp. 1–8.
12. Li, Z.; Xiong, F.; Lu, J.; Zhou, J.; Qian, Y. Material-Guided Siamese Fusion Network for Hyperspectral Object Tracking. In Proceedings of the ICASSP 2022—2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 2809–2813.
13. Lei, J.; Liu, P.; Xie, W.; Gao, L.; Li, Y.; Du, Q. Spatial–Spectral Cross-Correlation Embedded Dual-Transfer Network for Object Tracking Using Hyperspectral Videos. *Remote Sens.* **2022**, *14*, 3512. [CrossRef]
14. Li, Z.; Ye, X.; Xiong, F.; Lu, J.; Zhou, J.; Qian, Y. Spectral-Spatial-Temporal attention network for hyperspectral tracking. In Proceedings of the 2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Amsterdam, The Netherlands, 24–26 March 2021; pp. 1–5.
15. Bertinetto, L.; Valmadre, J.; Henriques, J.F.; Vedaldi, A.; Torr, P.H. Fully-convolutional siamese networks for object tracking. In *Computer Vision—ECCV 2016 Workshops, ECCV 2016*; Springer: Cham, Switzerland, 2016; pp. 850–865.
16. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High performance visual tracking with siamese region proposal network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8971–8980.
17. Zhang, Z.; Peng, H. Deeper and wider siamese networks for real-time visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4591–4600.
18. Li, B.; Wu, W.; Wang, Q.; Zhang, F.; Xing, J.; Yan, J. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019, Long Beach, CA, USA, 15–20 June 2019; pp. 4282–4291.
19. Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; Chen, S. SiamCAR: Siamese fully convolutional classification and regression for visual tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 6269–6277.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef] [PubMed]
21. Zhang, Z.; Peng, H.; Fu, J.; Li, B.; Hu, W. Ocean: Object-aware anchor-free tracking. In *Computer Vision—ECCV 2020, ECCV 2020*; Springer: Cham, Switzerland, 2020; pp. 771–787.

22. Xu, Y.; Wang, Z.; Li, Z.; Yuan, Y.; Yu, G. SiamFC++: Towards robust and accurate visual tracking with target estimation guidelines. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12549–12556.
23. Yu, Y.; Xiong, Y.; Huang, W.; Scott, M.R. Deformable siamese attention networks for visual object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 14–19 June 2020; pp. 6728–6737.
24. Shen, J.; Tang, X.; Dong, X.; Shao, L. Visual object tracking by hierarchical attention siamese network. *IEEE Trans. Cybern.* **2019**, *50*, 3068–3080. [CrossRef]
25. Gao, L.; Liu, P.; Ning, J.; Li, Y. Visual object tracking via non-local correlation attention learning. *Knowl.-Based Syst.* **2022**, *254*, 109666. [CrossRef]
26. Cai, H.; Zhang, X.; Lan, L.; Xu, L.; Shen, W.; Chen, J.; Leung, V.C. SiamATTRPN: Enhance Visual Tracking With Channel and Spatial Attention. *IEEE Trans. Comput. Soc. Syst.* **2023**, *11*, 1958–1966. [CrossRef]
27. Bao, H.; Shu, P.; Zhang, H.; Liu, X. Siamese-based twin attention network for visual tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 847–860. [CrossRef]
28. Fan, C.; Yu, H.; Huang, Y.; Shan, C.; Wang, L.; Li, C. SiamON: Siamese occlusion-aware network for visual tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *33*, 186–199. [CrossRef]
29. Zhang, H.; Liang, J.; Zhang, J.; Zhang, T.; Lin, Y.; Wang, Y. Attention-Driven Memory Network for Online Visual Tracking. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *35*, 17085–17098. [CrossRef]
30. Wang, Y.; Yan, L.; Feng, Z.; Xia, Y.; Xiao, B. Visual tracking using transformer with a combination of convolution and attention. *Image Vis. Comput.* **2023**, *137*, 104760. [CrossRef]
31. Van Nguyen, H.; Banerjee, A.; Chellappa, R. Tracking via object reflectance using a hyperspectral video camera. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition—Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 44–51.
32. Kiani Galoogahi, H.; Fagg, A.; Lucey, S. Learning background-aware correlation filters for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1135–1143.
33. Chen, L.; Zhao, Y.; Yao, J.; Chen, J.; Li, N.; Chan, J.C.W.; Kong, S.G. Object tracking in hyperspectral-oriented video with fast spatial-spectral features. *Remote Sens.* **2021**, *13*, 1922. [CrossRef]
34. Uzkent, B.; Rangnekar, A.; Hoffman, M.J. Tracking in aerial hyperspectral videos using deep kernelized correlation filters. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 449–461. [CrossRef]
35. Tang, Y.; Liu, Y.; Ji, L.; Huang, H. Robust Hyperspectral Object Tracking by Exploiting Background-Aware Spectral Information With Band Selection Network. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5.
36. Zhang, Y.; Li, X.; Wang, F.; Wei, B.; Li, L. A Fast Hyperspectral Object Tracking Method Based On Channel Selection Strategy. In Proceedings of the 2022 12th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Rome, Italy, 13–16 September 2022; pp. 1–5.
37. Li, Z.; Xiong, F.; Zhou, J.; Wang, J.; Lu, J.; Qian, Y. BAE-NET: A band attention aware ensemble network for hyperspectral object tracking. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; pp. 2106–2110.
38. Gao, L.; Chen, L.; Liu, P.; Jiang, Y.; Xie, W.; Li, Y. A Transformer-based Network for Hyperspectral Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11.
39. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, Utah, USA, 18-22 June 2018; pp. 7794–7803.
40. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer tracking. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Montreal, BC, Canada, 19–25 June 2021; pp. 8126–8135.
41. Chi, C.; Wei, F.; Hu, H. Relationnet++: Bridging visual representations for object detection via transformer decoder. *Adv. Neural. Inf. Process. Syst.* **2020**, *33*, 13564–13574.
42. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
43. Wang, N.; Zhou, W.; Wang, J.; Li, H. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 1571–1580.
44. Yan, B.; Peng, H.; Fu, J.; Wang, D.; Lu, H. Learning spatio-temporal transformer for visual tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 10448–10457.
45. Yu, B.; Tang, M.; Zheng, L.; Zhu, G.; Wang, J.; Feng, H.; Feng, X.; Lu, H. High-performance discriminative tracking with transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 9856–9865.

46. Zheng, Y.; Zhong, B.; Liang, Q.; Tang, Z.; Ji, R.; Li, X. Leveraging local and global cues for visual tracking via parallel interaction network. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *33*, 1671–1683.
47. Tang, C.; Wang, X.; Bai, Y.; Wu, Z.; Zhang, J.; Huang, Y. Learning spatial-frequency transformer for visual object tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 5102–5116.
48. Lin, L.; Fan, H.; Zhang, Z.; Xu, Y.; Ling, H. SwinTrack: A simple and strong baseline for transformer tracking. *Adv. Neural. Inf. Process. Syst.* **2022**, *35*, 16743–16754.
49. Cui, Y.; Jiang, C.; Wang, L.; Wu, G. Mixformer: End-to-end tracking with iterative mixed attention. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 13608–13618.
50. Zhao, C.; Liu, H.; Su, N.; Yan, Y. TFTN: A Transformer-Based Fusion Tracking Framework of Hyperspectral and RGB. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15.
51. Zhao, C.; Liu, H.; Su, N.; Wang, L.; Yan, Y. RANet: A reliability-guided aggregation network for hyperspectral and RGB fusion tracking. *Remote Sens.* **2022**, *14*, 2765. [CrossRef]
52. Zhao, C.; Liu, H.; Su, N.; Xu, C.; Yan, Y.; Feng, S. TMTNet: A Transformer-Based Multimodality Information Transfer Network for Hyperspectral Object Tracking. *Remote Sens.* **2023**, *15*, 1107. [CrossRef]
53. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inform. Process. Syst.* **2017**, *30*, 6000–6010.
54. Zhao, M.; Meng, Q.; Zhang, L.; Hu, X.; Bruzzone, L. Local and Long-Range Collaborative Learning for Remote Sensing Scene Classification. *IEEE Trans. Geosci. Remote. Sens.* **2023**, *61*, 1–15. [CrossRef]
55. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
56. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common objects in context. In *Computer Vision—ECCV 2014, ECCV 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755.
57. Fan, H.; Lin, L.; Yang, F.; Chu, P.; Deng, G.; Yu, S.; Bai, H.; Xu, Y.; Liao, C.; Ling, H. LaSOT: A high-quality benchmark for large-scale single object tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 5374–5383.
58. Huang, L.; Zhao, X.; Huang, K. GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 1562–1577. [CrossRef] [PubMed]
59. Muller, M.; Bibi, A.; Giancola, S.; Alsubaihi, S.; Ghanem, B. TrackingNet: A large-scale dataset and benchmark for object tracking in the wild. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 300–317.
60. Danelljan, M.; Häger, G.; Khan, F.S.; Felsberg, M. Discriminative Scale Space Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1561–1575. [CrossRef] [PubMed]
61. Danelljan, M.; Hager, G.; Shahbaz Khan, F.; Felsberg, M. Learning spatially regularized correlation filters for visual tracking. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 4310–4318.
62. Wang, N.; Zhou, W.; Tian, Q.; Hong, R.; Wang, M.; Li, H. Multi-cue correlation filters for robust visual tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Salt Lake City, UT, USA, 18–22 June 2018; pp. 4844–4853.
63. Chen, X.; Peng, H.; Wang, D.; Lu, H.; Hu, H. Seqtrack: Sequence to sequence learning for visual object tracking. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 14572–14581.
64. Wei, X.; Bai, Y.; Zheng, Y.; Shi, D.; Gong, Y. Autoregressive visual tracking. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 9697–9706.
65. Cai, W.; Liu, Q.; Wang, Y. HIPTrack: Visual Tracking with Historical Prompts. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 19258–19267.
66. Zhang, Z.; Qian, K.; Du, J.; Zhou, H. Multi-features integration based hyperspectral videos tracker. In Proceedings of the 2021 11th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS), Amsterdam, The Netherlands, 24–26 March 2021; pp. 1–5.
67. Tang, Y.; Huang, H.; Liu, Y.; Li, Y. A Siamese network-based tracking framework for hyperspectral video. *Neural Comput. Appl.* **2023**, *35*, 2381–2397. [CrossRef]
68. Wang, H.; Li, W.; Xia, X.G.; Du, Q.; Tian, J.; Shen, Q. Transformer-Based Band Regrouping With Feature Refinement for Hyperspectral Object Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–14. [CrossRef]

69. Chen, Y.; Tang, Y.; Su, X.; Li, J.; Xiao, Y.; He, J.; Yuan, Q. PHTrack: Prompting for Hyperspectral Video Tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–18. [CrossRef]
70. Chen, Y.; Yuan, Q.; Tang, Y.; Xiao, Y.; He, J.; Zhang, L. SPIRIT: Spectral awareness interaction network with dynamic template for hyperspectral object tracking. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [CrossRef]