remote sensing

MDPI

Article

TMTNet: A Transformer-Based Multimodality Information Transfer Network for Hyperspectral Object Tracking

Chunhui Zhao [1,2], Hongjiao Liu [1,2], Nan Su [1,2,*], Congan Xu [3,4], Yiming Yan [1,2] and Shou Feng [1,2]

[1] College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
[2] Key Laboratory of Advanced Marine Communication and Information Technology, Ministry of Industry and Information Technology, Harbin Engineering University, Harbin 150001, China
[3] Institute of Information Fusion, Naval Aviation University, Yantai 264000, China
[4] Advanced Technology Research Institute, Beijing Institute of Technology, Jinan 250300, China
* Correspondence: sunan08@hrbeu.edu.cn

Abstract: Hyperspectral video with spatial and spectral information has great potential to improve object tracking performance. However, the limited hyperspectral training samples hinder the development of hyperspectral object tracking. Since hyperspectral data has multiple bands, from which any three bands can be extracted to form pseudocolor images, we propose a Transformer-based multimodality information transfer network (TMTNet), aiming to improve the tracking performance by efficiently transferring the information of multimodality data composed of RGB and hyperspectral in the hyperspectral tracking process. The multimodality information needed to be transferred mainly includes the RGB and hyperspectral multimodality fusion information and the RGB modality information. Specifically, we construct two subnetworks to transfer the multimodality fusion information and the robust RGB visual information, respectively. Among them, the multimodality fusion information transfer subnetwork is designed based on the dual Siamese branch structure. The subnetwork employs the pretrained RGB tracking model as the RGB branch to guide the training of the hyperspectral branch with little training samples. The RGB modality information transfer subnetwork is designed based on a pretrained RGB tracking model with good performance to improve the tracking network's generalization and accuracy in unknown complex scenes. In addition, we design an information interaction module based on Transformer in the multimodality fusion information transfer subnetwork. The module can fuse multimodality information by capturing the potential interaction between different modalities. We also add a spatial optimization module to TMTNet, which further optimizes the object position predicted by the subject network by fully retaining and utilizing detailed spatial information. Experimental results on the only available hyperspectral tracking benchmark dataset show that the proposed TMTNet tracker outperforms the advanced trackers, demonstrating the effectiveness of this method.

Keywords: hyperspectral object tracking; Transformer; multimodality; Siamese network

Citation: 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. https://doi.org/10.3390/rs15041107

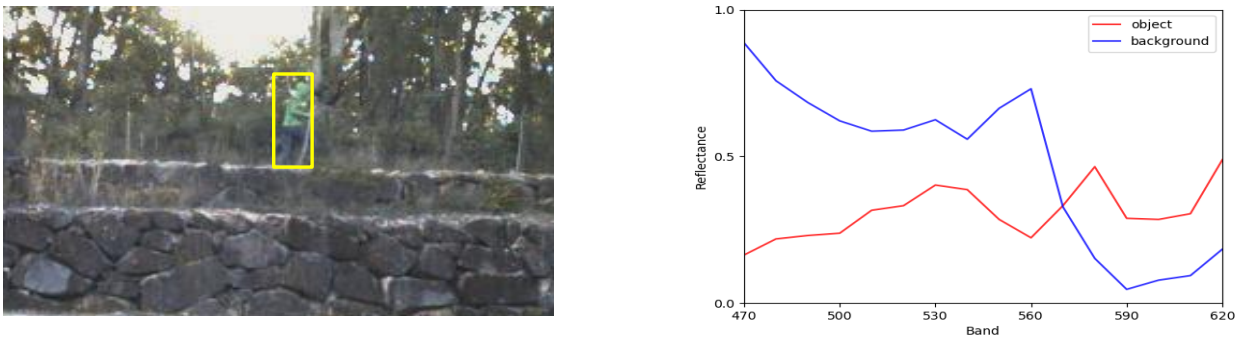
Academic Editors: Yanfei Zhong, Pedram Ghamisi, Jun Zhou, Jocelyn Chanussot and Fengchao Xiong

Received: 15 January 2023
Revised: 6 February 2023
Accepted: 15 February 2023
Published: 17 February 2023

1. Introduction

Hyperspectral object tracking is a challenging task emerging recently [1–3], which can be applied in video surveillance camouflage targets, autonomous driving, and so on. Its purpose is to estimate the object's state (e.g., position, size, etc.) in subsequent frames by that of the object in the initial frame in the hyperspectral video. Currently, most tracking algorithms are developed for RGB video research and have made some achievements [4–6]. However, the RGB modality image has inherent limitations in describing the physical characteristics of objects, making it easy to cause RGB-based tracker drifts in some complex but common scenarios, such as the object and backgrounds' colors being similar. Compared with the RGB image that describes visual information only by red, green, and blue channels, the hyperspectral image (HSI) with a three-dimensional structure can record the location

of the object space and the continuous spectral information simultaneously. As shown in Figure 1, HSI can provide additional spectral information to break through the limitations of visual characteristics, which proves that HSI has the potential to cope with the challenges in the tracking process. Therefore, using hyperspectral video to perform the tracking task can offer more opportunities for achieving high-performance tracking, which has significant research value.



(a) **(b)**

Figure 1. An example of the object and background colors are similar in the RGB image, but their spectral information is inconsistent. (**a**) shows the relationship between the object and the background in the RGB image, in which the object is marked with a yellow bounding box. (**b**) displays the spectral response curves of two pixels from the object and the background, respectively.

Some works have preliminary explored hyperspectral object tracking methods [1,7–10] in recent years. Similar to the RGB object tracking method, the hyperspectral object tracking algorithm can be divided into two kinds; one is based on correlation filtering, and the other is based on deep learning (DL) [11,12]. The MHT [1] method proposed by Xiong et al. is a representative correlation filtering-based hyperspectral object tracking method. MHT adopts two feature descriptors to characterize material information of HSIs and further embeds them into the background-aware correlation filter, yielding the tracking based on material. However, compared with the deep features obtained by deep neural networks, the handcrafted features usually adopted by the correlation filtering method have difficulty with fully describing hyperspectral information, which often limits the hyperspectral object tracking performance. Therefore, applying the DL method in the hyperspectral object tracking field is more competitive for accurately predicting the object's state in the tracking process.

However, the limited amount of hyperspectral image sequences cannot meet the requirements of deep learning for large-scale training samples, which undoubtedly makes it difficult to promote the development of DL-based hyperspectral tracking algorithms [13,14]. Compared with HSI sequences, RGB image sequences have massive labeled samples and richer visual details (such as texture, color, and so on). Thus, the RGB object tracking method based on DL often has higher tracking accuracy. Therefore, exploring how to transfer the advantages of the DL-based RGB modality tracking method to hyperspectral tracking to alleviate the problem of low model accuracy and insufficient generalization ability caused by the shortage of training sample data in hyperspectral tracking is crucial for effectively using the DL method to improve the performance of hyperspectral modality object tracking.

At present, the method of successfully transferring the advantages of the RGB modality tracking method based on DL to the field of hyperspectral object tracking is to process hyperspectral modality data using the RGB tracking model based on DL trained by large-scale datasets to capture robust visual-similar features from the hyperspectral modality. These methods improve tracking performance by successfully transferring the robust RGB modality information in the hyperspectral object tracking process [2,3,15]. The BAE-Net [2] method proposed by Li et al. is an excellent and representative DL-based work. BAE-Net

first introduces a band attention module to learn the relationship among hyperspectral bands for generating band weights and divides the hyperspectral image into multiple three-channel images according to these weights. Then, these images are input into a deep RGB tracking model, transferring multiple visual-similar information from hyperspectral data for the integrated prediction of the object position. Consistent with the idea of BAE-Net, the SST-Net [3] method proposed by Li et al. also divides HSI bands and uses the depth tracker for integrated tracking. The difference is that SST-Net considers the spatial–spectral–temporal information in the hyperspectral video when acquiring the importance of bands, which can model the relationship between bands of HSIs better, thus converting HSIs into more valuable three-band images for depth tracking. Unlike the above methods, the HA-Net [15] method proposed by Liu et al. is another meaningful and representative work of the DL-based hyperspectral object tracking task. HA-Net leverages the dual Siamese network framework to perform hyperspectral object tracking, using the hyperspectral information to improve the performance of the RGB Siamese tracking network, which can make the model more discriminative. Specifically, the RGB Siamese network is used to obtain visual-similar features from false-color images converted from hyperspectral data and then get classification and regression response maps of the false-color data. The hyperspectral Siamese network is used to obtain the classification response map of the hyperspectral data. Two classification response maps are merged to enhance the network's ability to distinguish the object and the background. Unfortunately, although they have achieved preliminary success in transferring the RGB tracking advantages to hyperspectral tracking by using the DL-based RGB tracking model to transfer the RGB modality information, they still do not fully play the role of hyperspectral information to improve object tracking performance.

Effective use of the pretrained RGB tracking model based on DL to transfer RGB modality information in hyperspectral object tracking while fully using hyperspectral data information is essential to achieve high-performance hyperspectral tracking. Multimodality fusion tracking tasks have become popular recently [16–18], which can improve tracking performance by efficiently combining the information of different modalities to supplement the inherent defects of single-modality. It is well known that extracting any three bands from hyperspectral data can form pseudocolor images. Therefore, the hyperspectral object tracking task can be regarded as multimodality object tracking based on the hyperspectral and pseudocolor video. Thus, while using the pre-trained RGB model to transfer RGB modality information, it is worth to explore that introducing the idea of multimodality tracking into the object tracking field based on the single hyperspectral modality, which can realize the full utilization of hyperspectral data by effectively transferring the fusion information of multimodality data composed of RGB and hyperspectral, thereby improving the performance of hyperspectral tracking. In addition, the successful application of the Transformer model in multimodality tasks [19–21] shows that the model can achieve the purpose of information combination by efficiently capturing different modality relations to fuse information. Therefore, it has great potential to improve the performance of the tracking task by using the Transformer model to combine different modality information.

Based on these have been mentioned, we propose a Transformer-based multimodality information transfer network (TMTNet) for hyperspectral object tracking, aiming to fully transfer the information of multimodality data composed of RGB data and hyperspectral data to enhance the object tracking's performance based on single hyperspectral modality. In this work, the multimodality information that needs to be transferred mainly includes the fusion information of multimodality data composed of RGB and hyperspectral and the RGB modality information. The information transfer is realized through the corresponding pretrained network to alleviate the deep model's low accuracy and insufficient generalization ability caused by the lack of hyperspectral training samples. The RGB pretrained network is trained through tens of millions of RGB training samples, which can predict the object location robustly in unknown scenes. However, relative to the RGB data scale, no large-scale dataset containing RGB and hyperspectral video data pairs can be used to provide the training samples required for the pretrained multimodality fusion network. To

this end, we adopt the dual branch fusion structure, which uses the DL-based pretrained RGB model as the RGB branch to process RGB data and uses the RGB branch to guide the training of the hyperspectral branch to realize that modeling the general representation ability of hyperspectral features with a small number of training samples, thus obtaining the pretrained RGB-hyperspectral multimodality fusion model with certain generalization ability. It is worth noting that the existing combination of RGB and hyperspectral video data is not entirely ideal (it has some differences, such as a spatial resolution difference), but this does not affect the construction of the relation of RGB and hyperspectral video data using the Siamese network based on the known two modality ground truth. This is because, in the training process, the template patch and the search region as the actual input of the Siamese network are all clipped based on the ground truth of each modality data, and the size of the corresponding area after the clipping of the two modality data is fixed and the same. Therefore, even if the two modality data are not entirely matched, it has little effect on the Siamese network-based fusion model for training the two modalities.

It is well known that multimodality fusion information not only contains the advantages of each modality data but also complements the shortcomings of single-modality data, which is conducive to improving tracking performance. To fully utilize hyperspectral information from the perspective of multimodality fusion information transfer, we construct a multimodality fusion information transfer subnetwork (trained by the multimodality data composed of RGB and hyperspectral) in TMTNet, to predict the object position in the hyperspectral video by capturing the multimodality-similar fusion information from hyperspectral data in the testing process. The critical parts of the subnetwork include a dual Siamese network-based branch structure and a multimodality fusion module, which are used to process different modality data and fuse their semantic information, respectively. Specifically, a pretrained RGB Siamese network model based on DL is used as the RGB branch to process pseudocolor data to obtain general, robust, and descriptive visual-similar features. Then, a Siamese 3D CNN is designed as the hyperspectral branch to process hyperspectral data. The Siamese 3D CNN obtains the hyperspectral modality-specific information by adopting the 3D convolution kernel to slide jointly between the spatial and spectral dimensions of the hyperspectral data. In addition, given the Transformer model's advantage in combining multimodality information, the multimodality fusion module is designed based on the Transformer model. This module (termed TIIM) adopts the self-attention mechanism of the Transformer to interact the semantic information generated by different modality branches adaptively to achieve multimodality information fusion. Therefore, the constructed multimodality fusion information transfer subnetwork can obtain multimodality-similar fusion information from hyperspectral data by effectively combining pseudocolor and hyperspectral information based on ensuring a certain generalization ability to achieve accurate prediction of the object location.

To further improve the tracking network's generalization and accuracy, on the basis of the multimodality fusion information transfer subnetwork, we introduce a good-performance RGB tracking model as the other tracking subnetwork into TMTNet, for transferring the robust RGB modality information. The RGB modality information transfer subnetwork maximizes the ability of the network to track objects in unknown complex scenes by adding robust visual-similar features of the pseudocolor data to the tracking model. Then, two sets of response maps generated by two subnetworks are employed to jointly predict the object's position to make the tracking results more accurate. The mentioned above are essential components of the subject network in TMTNet. In addition, to obtain a higher-quality estimation bounding box of object tracking, we also add a spatial optimization module (SOM) to TMTNet, which further optimizes the object position predicted by the subject network by fully retaining and utilizing detailed spatial information. The experimental results on the only available hyperspectral tracking benchmark dataset currently [1] show that our method achieves leading performance, outperforming advanced trackers. The proposed TMTNet is an extension of our previous work TrTSN [22], in which TrTSN is the champion scheme of the Hyperspectral Object Tracking Competition 2022.

Compared with TrTSN, TMTNet employs the independent RGB tracking model trained by large-scale datasets as the RGB modality information transfer subnetwork and adds a spatial optimization module to optimize the tracking performance, achieving a similar tracking accuracy to that of TrTSN, which indicates that the hyperspectral object tracking method designed from the perspective of multimodality information transfer is flexible, simple, and effective. The main contributions of this paper are summarized as follows.

1. We propose a multimodality information transfer network for hyperspectral object tracking, which improves the tracking performance based on the single hyperspectral modality by efficiently transferring the information of multimodality data composed of RGB and hyperspectral. This is the first time that the idea of multimodality tracking is introduced into single-modality object tracking, which provides a new idea for achieving high-performance hyperspectral object tracking.
2. We construct two subnetworks in the subject network of TMTNet to transfer the semantic information of multimodality data from different angles in the hyperspectral tracking process, thus improving the network's ability to predict the object's location. Among them, one subnetwork is used to improve the tracking performance by transferring the multimodality fusion information containing the complementary features of RGB and hyperspectral data. The other subnetwork is used to enhance the tracking network's generalization and accuracy by transferring robust RGB visual features using the deep-learning-based RGB model trained by large-scale datasets.
3. We design an information interaction module based on Transformer (TIIM) in the multimodality fusion subnetwork of the subject network, which uses the Transformer's self-attention mechanism to adaptively capture the potential interactions between the semantic information generated by different modality branches to achieve multimodality information fusion. As far as we know, this is the first application of the Transformer model to combine different semantic information in hyperspectral object tracking.

The rest of this paper is organized as follows. In Section 2, we describe the Transformer-based multimodality information transfer network in detail. The experimental detail is presented in Section 3. In Section 4, we present the experimental results and analysis, and finally, in Section 5, we conclude the paper.

2. Methods

2.1. Network Architecture

The proposed Transformer-based multimodality information transfer hyperspectral object tracking network (TMTNet) transfers the information of multimodality data composed of RGB and hyperspectral to hyperspectral tracking by using the corresponding network model, which can fully use hyperspectral information from different angles to achieve accurate prediction of object location. The network not only contains a subject network part to predict the object's primary location but also a spatial optimization module (SOM) to optimize the quality of the object bounding box. The subject network contains a multimodality fusion information transfer subnetwork and an RGB modality information transfer subnetwork, which are used to obtain multimodality-similar fusion information and visual-similar information from hyperspectral data, respectively, aiming to achieve the tracking performance improvement by fully using hyperspectral data. In addition, this network has an anchor-free architecture, making the tracking network more concise. The architecture of the TMTNet is introduced in Figure 2.

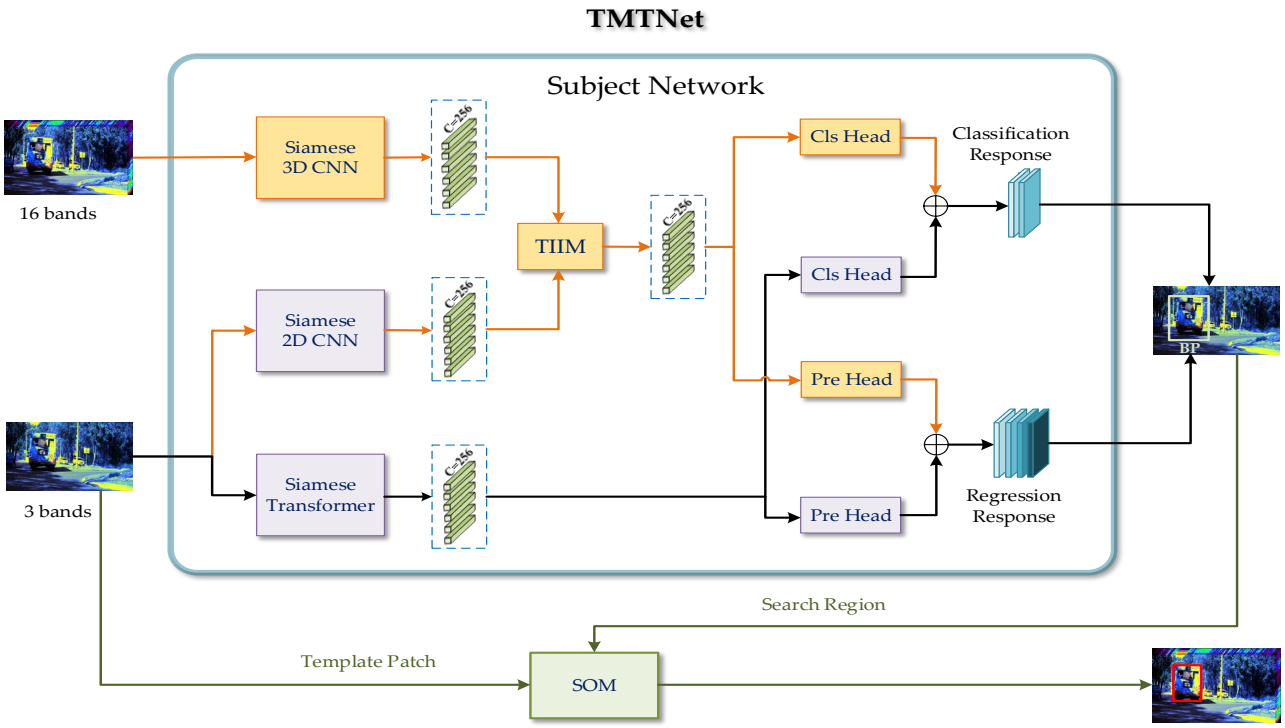


Figure 2. The proposed TMTNet's architecture. The part with the blue border box describes the structure of the subject network. The modules connected by orange arrows in the subject network are components of the multimodality fusion information transfer subnetwork. The remaining modules are components of the RGB modality information transfer subnetwork. The yellow module is related to processing hyperspectral data, whereas the purple module is related to processing pseudocolor data. TIIM is the Transformer-based information interaction module, $\oplus$ represents the merge operator and SOM describes the spatial optimization module. 'BP' represents the object's bounding box predicted by the subject tracker.

In this work, the hyperspectral data are regarded as the multimodality data composed of the hyperspectral data (with 16 bands) and pseudocolor data (consisting of 3-band hyperspectral data). As we can see, the subject network in TMTNet mainly includes three Siamese network branches, an information interaction module based on Transformer (TIIM), and two sets of prediction heads. Each set of prediction heads consists of a classification prediction head and a regression prediction head. Among them, the Siamese 3D CNN branch, the Siamese 2D CNN branch, the TIIM, and a set of prediction heads are components in the multimodality fusion information transfer subnetwork. The rest parts belong to the RGB modality information transfer subnetwork. The overall input of the network is the hyperspectral data and pseudocolor data formed by the hyperspectral data. First, three Siamese network branches are adapted to process the hyperspectral and pseudocolor data to generate three different semantic information. The Siamese 3D CNN branch is used to process hyperspectral data, while the other two are applied to process pseudocolor data. Second, the TIIM is adopted to integrate the semantic information obtained by Siamese 3D CNN and Siamese 2D CNN branches adaptively to generate the multimodality-similar fusion feature that includes the information of the hyperspectral data and pseudocolor data. Finally, two sets of prediction heads are used to predict the multimodality-similar fusion feature obtained by the second step and the visual-similar feature obtained from the Siamese Transformer branch. The response-level fusion method is used to merge the generated two sets of response maps to obtain the final response maps. The final classification and regression response maps are employed to jointly predict the object's primary location. TMTNet also contains a spatial optimization module, which is used to optimize the object's

primary location predicted by the subject tracker, thereby achieving higher-performance object tracking.

2.2. The Subject Network of TMTNet

The subject network is vital to ensure the tracking accuracy of TMTNet. The subject network consists of a multimodality fusion information transfer subnetwork and an RGB modality information transfer subnetwork. From the perspective of multimodality fusion information transfer, the multimodality fusion information transfer subnetwork obtains multimodality-similar fusion information of hyperspectral data to improve tracking performance. From the standpoint of RGB modality information transfer, the RGB modality information transfer subnetwork gets robust visual-similar features of hyperspectral data to improve the ability of the network to predict the object position in unknown complex scenes accurately. Specifically, the multimodality fusion information transfer subnetwork includes two Siamese network branches, a TIIM, and a set of prediction heads. The RGB modality information transfer subnetwork has a Siamese network branch and another set of prediction heads. The hyperspectral video data is processed by two subnetworks and generates two response map sets. Then, the response-level fusion method is used to merge them as the final response maps for predicting the object position. The details are described as follows.

2.2.1. Three Siamese Network Branches

Fully obtaining the hyperspectral semantic information is the basis for enhancing the network's ability to accurately predict the object's location in the hyperspectral tracking process. Given the Siamese trackers' exemplary performance in RGB object tracking [23–26], we employ the Siamese network to extract hyperspectral data features. We construct three Siamese network branches (Siamese 3D CNN, Siamese 2D CNN, and Siamese Transformer) to fully get the hyperspectral semantic information from different angles. The hyperspectral data is first regarded as the multimodality data composed of the hyperspectral data (with 16 bands) and pseudocolor data (consisting of 3-band hyperspectral data) and then input into the network.

Hyperspectral data and pseudocolor data need to be preprocessed before inputting Siamese network branches. Generally, the first frame of the video data containing the object ground truth is selected as the template image, and the rest of the frames are the search images. In the template image, the region extending from the object's center to twice the side length is viewed as the template patch, which contains information about the object and its local surrounding scene. In the current frame, the search region is the area that extends from the object center in the previous search image to four times the length of the side. The search region typically covers the object's possible range. The template patch and search region are then sent to the Siamese branch for processing.

Each Siamese network branch has the backbone and information transmission parts. The backbone is applied to extract the template patch and search region features. The information transmission part is utilized to transmit the template information to the search region. The Siamese network's structure is shown in Figure 3. Each Siamese network has two backbones with shared parameters and the same structure. The structure or parameters of the backbone in the three Siamese branches are inconsistent.

In the multimodality fusion information transfer subnetwork, inspired by [27], we design the 3D convolution neural network as the backbone in the Siamese 3D CNN branch. The spatial–spectral joint information of hyperspectral data can be extracted by utilizing the 3D convolution kernel naturally and elegantly, as shown in Figure 4. The kernel size in the backbone of the Siamese 3D CNN branch is listed in Table 1. In addition, in the Siamese 2D CNN branch, the pretrained ResNet-50 is exploited in [28] as the backbone to obtain the visual-similar feature of pseudocolor data.

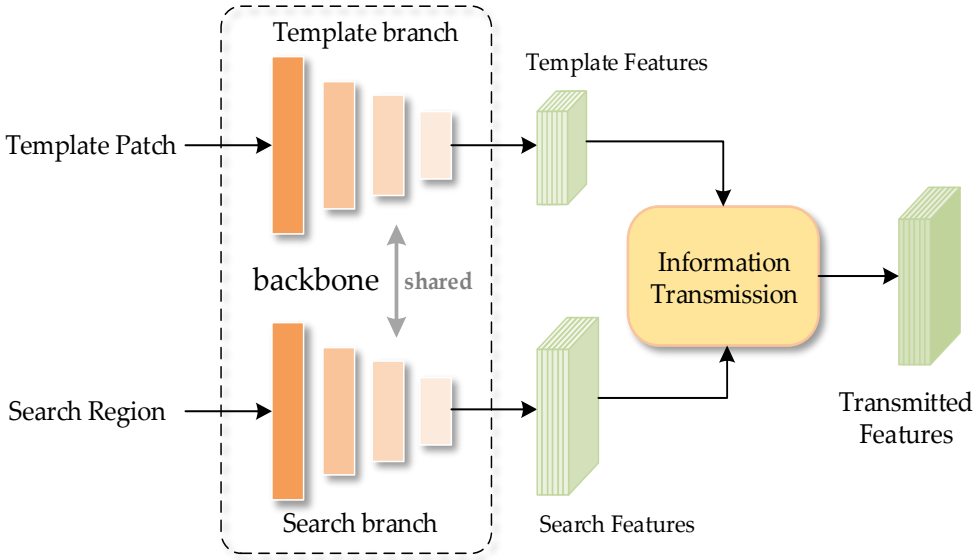


Figure 3. The Siamese network's structure. The Siamese network includes the backbone and information transmission parts.

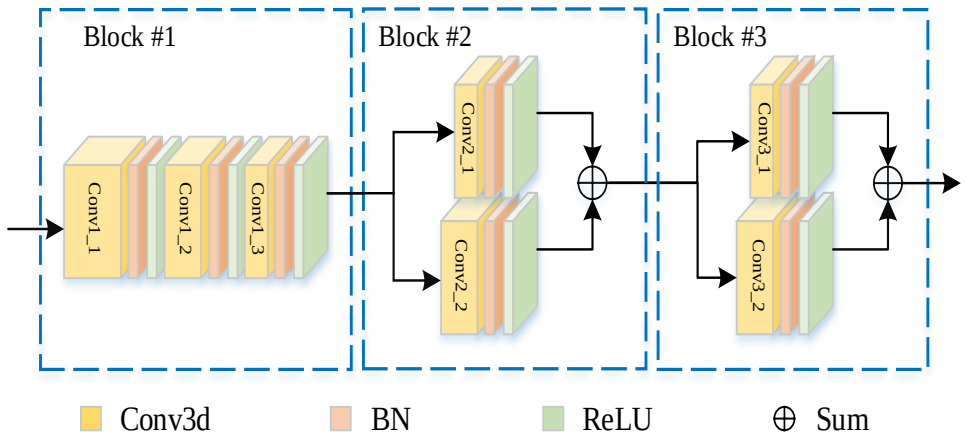


Figure 4. The backbone's structure in the Siamese 3D CNN branch.

Table 1. The convolutional layers' parameter in the Siamese 3D CNN's backbone.

Block Name	Kernel Name	Kernel Number	Kernel Size (H,W,B)
Block #1	conv1	256	7,7,5
	conv1_1	256	3,3,3
	conv1_2	256	3,3,1
Block #2	conv2_1	256	1,1,1
	conv2_2	256	1,1,3
Block #3	conv3_1	256	1,1,1
	conv3_2	256	1,1,3

Kernel size represents the kernel size in the backbone of the Siamese 3D CNN branch.

In this subnetwork, the cross-correlation operation is adopted to transmit the template patch information and the search region information in Siamese 3D CNN and Siamese 2D CNN branches. Notably, the Siamese 3D CNN branch adopts two cross-correlation operations to calculate the depth-correlation of features obtained by Block #2 and Block #3

of the HSI backbone. In addition, the Siamese 2D CNN branch uses three cross-correlation operations to perform depth-correlation calculations of the RGB backbone's features. A total of two depthwise cross-correlation features (transmitted features) are generated by Siamese 3D CNN and Siamese 2D CNN branches, which need to be further input into the TIIM to fuse different modality information.

In the RGB modality information transfer subnetwork, the pretrained ResNet-50 is also employed as the backbone of the Siamese Transformer branch to process pseudocolor images. In addition, this branch introduces the Transformer's attention module into the information transmission part (termed TIT), which can fully transmit the information of pseudocolor data by considering the nonlinear interaction between the global information of the template patch and the search region. TIT is the significant component of the Siamese Transformer branch, composed of four feature transmission layers and a separate feature transmission part. The structure of TIT is shown in Figure 5. Each feature transmission layer includes two Feature Self-Augment (FSA) modules and two Feature Cross-Augment (FCA) modules. The FSA module is used to enhance the template patch and search region's features, and the FCA module plays the role of transmitting both pieces of information. Spatial position coding adds position information to the FSA and FCA modules. The FSA module and the FCA module's structure are shown in Figure 6.

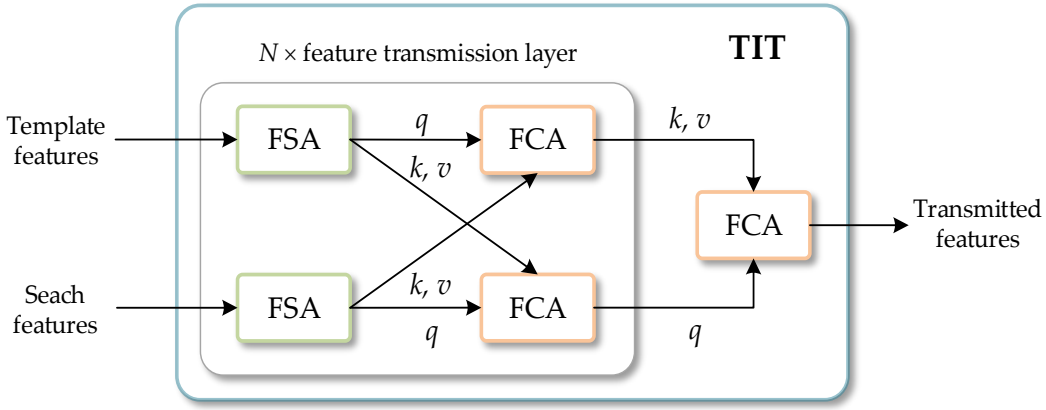


Figure 5. The structure of TIT. FSA represents the Feature Self-Augment module, and FCA means the Feature Cross-Augment module.

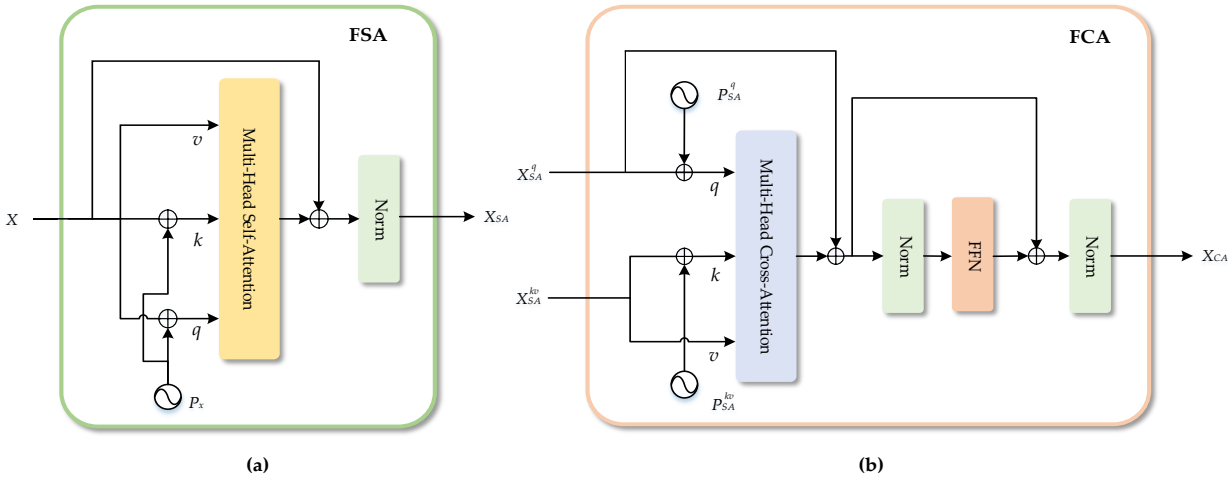


Figure 6. The structure of TIT. (**a**) shows the FSA structure, and (**b**) displays the FCA structure.

From Figure 6a, the FSA module has one input and one output. In the FSA module, the features are enhanced using the multiheaded self-attention with the residual form. This

module achieves image feature enhancement by better associating the semantic information of the image, which can be described as

$$X_{SA} = X + Multi\_Head(X + P_x, X + P_x, X), \tag{1}$$

the symbol $P_x \in \mathbb{R}^{HW \times C}$ indicates the spatial position coding, and $X_{SA} \in \mathbb{R}^{HW \times C}$ represents the enhanced features.

Figure 6b shows the FCA module has two inputs and one output. The features of the template patch and the search region are enhanced by the FSA module and then used as the input of the FCA module, which can use the multihead cross-attention in the FCA module to achieve the object information transmission better. In addition, a Feedforward Network (FFN) is added to the FCA module to increase the model's fitting ability. The FCA module can be described as

$$\tilde{X}_{CA} = X^q_{SA} + Multi\_Head(X^q_{SA} + P^q_{SA}, X^{kv}_{SA} + P^{kv}_{SA}, X^{kv}_{SA}), \tag{2}$$

$$X_{CA} = \tilde{X}_{CA} + FFN(\tilde{X}_{CA}). \tag{3}$$

The symbol $X^q_{SA} \in \mathbb{R}^{H_1W_1 \times C}$ is one branch's input feature, and $X^{kv}_{SA} \in \mathbb{R}^{H_2W_2 \times C}$ stands for that of the other. Correspondingly, $P^q_{SA} \in \mathbb{R}^{H_1W_1 \times C}$ is the spatial position coding of X^q_{SA}, and $P^{kv}_{SA} \in \mathbb{R}^{H_2W_2 \times C}$ is that of X^{kv}_{SA}. $X_{CA} \in \mathbb{R}^{H_1W_1 \times C}$ represents the output of the FCA module.

More details can be found in the literature [29].

2.2.2. Transformer-Based Information Interaction Module

The Transformer model [30] is constructed based on the attention mechanism, which makes a good performance in multimodality fields, such as image–text conversion [31], video retrieval [32], and multimodality detection [33]. Therefore, the Transformer model has great potential in capturing the relationship between different modality information. Therefore, we design an information interaction module based on Transformer (TIIM) to fuse multimodality information. The module utilizes the Transformer's self-attention mechanism to adaptively capture the potential interactions between the semantic information obtained from Siamese 3D CNN and Siamese 2D CNN branches to achieve multimodality information fusion. The structure of TIIM is shown in Figure 7.

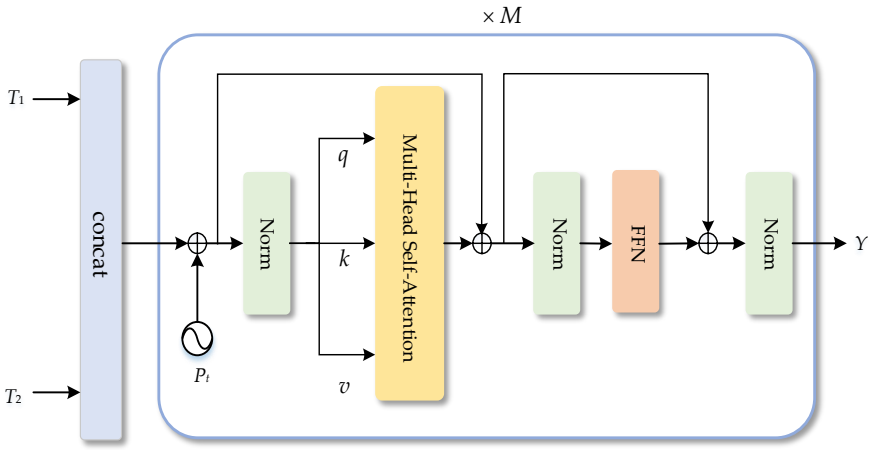


Figure 7. The structure of TIIM.

First, two features obtained by different Siamese branches, $T_1 \in \mathbb{R}^{H_TW_T \times C}$ and $T_2 \in \mathbb{R}^{H_TW_T \times C}$, are concatenated to get TIIM's input $T \in \mathbb{R}^{2H_TW_T \times C}$:

$$T = Concat(T_1, T_2). \tag{4}$$

Then, the input information is adaptively and fully integrated by the mechanism of multihead self-attention with the residual's form:

$$\tilde{T} = T + Multi\_Head(T + P_t, T + P_t, T), \tag{5}$$

where $P_t \in \mathbb{R}^{2H_T W_T \times C}$ encodes the spatial position of T.

In addition, an FFN module is used for this module, and finally, the output Y can be described as

$$Y = \tilde{T} + FFN(\tilde{T}). \tag{6}$$

2.2.3. Response-Level Fusion

Like most anchor-free Siamese trackers, the proposed TMTNet tracker uses classification and regression response maps to predict the object's location. Two sets of prediction heads (each set of prediction heads includes a classification prediction head and a regression prediction head) are used to process the multimodality-similar fusion feature of hyperspectral data generated by TIIM and the visual-similar feature of hyperspectral data obtained by the Siamese Transformer branch, respectively, to get two sets of response maps. We adopt the response-level fusion method to integrate two sets of response maps into a set of average response maps and use the merged response maps to predict the object in the hyperspectral tracking process. The final response maps R is shown as follows:

$$R = \frac{1}{N} \sum_{i=1}^{N} R_i, \tag{7}$$

where N represents the total number of interactive features, and R_i represents the response map of the ith interactive feature.

Compared with the decision-level fusion method that needs to directly integrate the final prediction results (the object bounding box predicted by the sub-network) of the two sub-networks, the classification and regression maps of the transferred multi-modality features of two sub-networks are fused at the response-level, which not only reduces the excessive dependence on the prediction results but also uses the information of different transferred features effectively, improving the tracking network's performance.

2.3. The Spatial Optimization Module

Inspired by [34], to further obtain a higher-quality estimation bounding box of object tracking, a spatial optimization module (SOM) is introduced in the tracking framework, which further optimizes the object position predicted by the subject tracker by fully retaining and utilizing detailed spatial information, thereby achieving higher performance object tracking.

The SOM's structure is also designed based on the Siamese network, as shown in Figure 8. Unlike the information transmission part mentioned above, the module utilizes pixelwise correlation operations to transmit features for preserving spatial detail information better. In addition, to fully use spatial information, the module adopts the corner prediction head and the auxiliary mask prediction head to predict object position for obtaining a more accurate object bounding box.

Specifically, the template branch of the SOM is initialized in the same way as the subject tracker, which is initialized by the template frame with ground truth. In each subsequent frame, the search branch of SOM predicts the object position further based on the concentric search region extended twice of the object bounding box indicated by the subject tracker, to obtain a more accurate object bounding box.

It can be noted that the SOM's search region is about twice the object's size, which is smaller than that of the subject tracker. There are two main reasons for choosing a smaller search region. One reason is that a smaller search region suppresses cluttered backgrounds and enables the model to be more concerned with detailed spatial information, facilitating precise positioning. The other reason is that the smaller search region also reduces the

computational cost so that the optimization module can improve the tracking performance of the subject tracker with almost no speed loss.

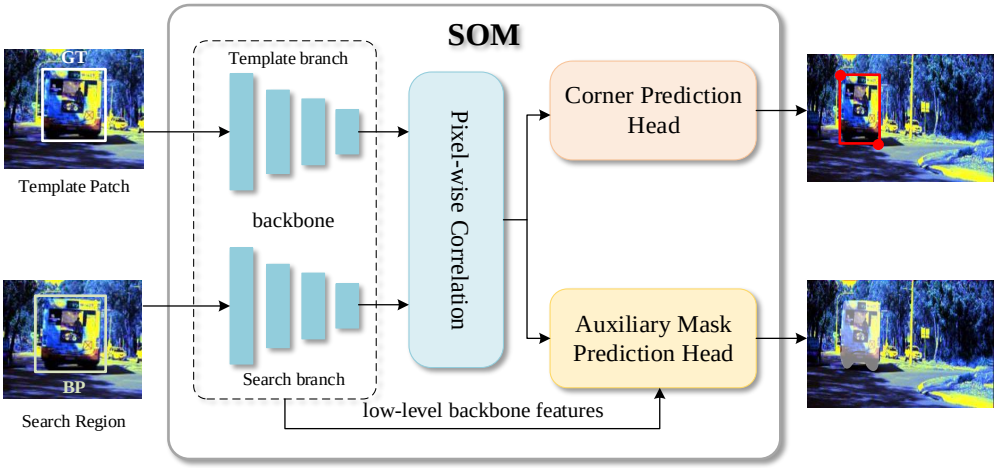


Figure 8. The structure of the SOM. Among them, 'GT' represents the groundtruth of the object, and 'BP' represents the object's bounding box predicted by the subject tracker.

2.3.1. Pixelwise Correlation

For the SOM with the Siamese structure, preserving the spatial detail information in the information transmission part between the template patch and the search region as much as possible is critical for optimizing the tracking results effectively. Most methods with the Siamese structure utilize single cross-correlation [35] or deep cross-correlation [28,36] operations for information transmission at present. However, the naive correlation operator or the depth correlation operator uses the entire template patch feature as the kernel of the search region feature to calculate the correlation and transmit information, which blurs the spatial information to some extent. Therefore, information transmission should be carried out in a way that is more beneficial to preserve spatial details in SOM.

In this work, SOM adopts the pixelwise correlation [37] operation to transmit the template patch and search region's information, to form feature representations with rich spatial detail information. The schematic diagram of pixel-level correlation operation is shown in Figure 9. Pixelwise correlation is used to achieve information transmission between pixels in the template patch and search region. Denote the template patch and search region's features extracted from the optimization module's backbone as $F_t \in \mathbb{R}^{C\times H_t\times W_t}$ and $F_s \in \mathbb{R}^{C\times H_s\times W_s}$, respectively. Among them, C is the feature channels' number, H_t (W_t) and H_s (W_s) are the height (width) of the template patch and the search region's feature map. To calculate the pixelwise correlation, first, the template patch features are divided into $H_t \times W_t$ small kernels $F_{ti} \in \mathbb{R}^{C\times 1\times 1}$, and the template patch features set can be expressed as $\mathcal{F}_t = \{F_{ti} \mid i = 1, 2, ..., H_t \times W_t\}$. After that, the correlation between each element F_{ti} in the template feature set $\mathcal{F}_t$ and the search region feature F_s is calculated separately. After correlation, $H_t \times W_t$ correlation maps $C_i \in \mathbb{R}^{H\times W}$ with the size of $H \times W$ can be obtained, and the set of correlation maps can be denoted as $\mathcal{C} = \{C_i \mid i = 1, 2, ..., H_t \times W_t\}$. The process can be described as follows:

$$\mathcal{C} = \{C_i | C_i = F_{ti} \star F_s\}_{i=1,2,...,H_t\times W_t}, \tag{8}$$

where $\star$ represents the naive correlation operator.

Pixelwise correlation ensures that each pixel in the template frame feature is associated with a correlation map, which can fully preserve the spatial detail information of the object and avoid the feature blurring caused by the large correlation window that results in insufficient utilization of spatial information. Therefore, using the pixelwise correlation operation in the information transmission part to transmit information is beneficial for further optimizing the object position predicted by the subject tracker.

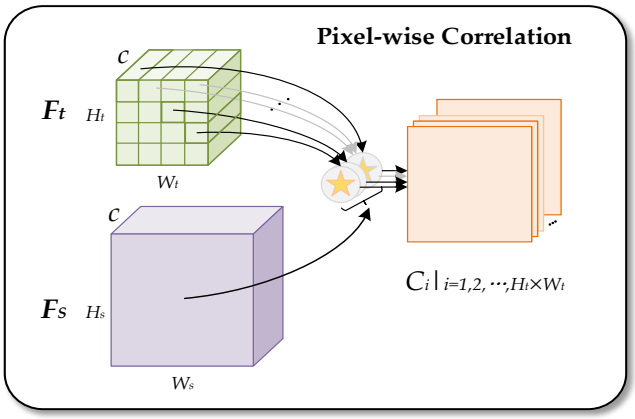


Figure 9. The schematic diagram of pixel-level correlation operation. The yellow star symbol denotes the naive correlation operator.

2.3.2. Corner Prediction Head

In the optimization module, selecting the prediction head that can fully use spatial information to estimate the object bounding box is important for successfully optimizing the object position predicted by the subject tracker. Many deep-learning-based Siamese trackers [28,35] employ a two-stage strategy to predict the current frame object state. Generally, the two-stage strategy is achieved by two prediction heads. First, a prediction head is used to locate the object roughly, and then the other head is utilized to refine results from the previous coarsely position. However, the use of the optimization module is under the condition that the primary position of the object is known (which is obtained by the subject tracker). Therefore, the prediction head required by the optimization module does not need to have the function of coarse positioning but needs to have a higher fine prediction function.

There are two common Siamese tracking refinement prediction heads: the RPN style refinement prediction head and the RCNN style refinement prediction head. The RPN style refinement prediction head mainly uses each feature point in the feature map to predict the four-dimensional coordinates of the bounding box. Each feature point encodes spatial information into the channel, so a single feature point can be used to predict the object boundary box. However, the spatial information of the object described by the feature points at different positions is inconsistent. The RPN-style method does not consider the relationship between the feature points at different positions, ignoring the information in the spatial distribution of the feature map. Therefore, the RPN-style method is not conducive to improving the prediction accuracy of the object's bounding box. The RCNN style refinement prediction head converts the feature map into the feature vector, then uses the fully connected layer to estimate the object's bounding box. Although this method utilizes the whole feature map to predict the object's position, it will destroy the spatial information when the feature map is transformed. Thus, the RCNN style refinement prediction head is unsuitable for optimizing the object boundary.

Compared to refinement prediction methods that have been mentioned (direct regression box coordinates), predicting two corners of an object from two heat maps is more competitive for refining the object's spatial position [34]. Therefore, SOM adopts a corner prediction head to predict the object's top-left corner and the bottom-right corner for obtaining the object's rectangular bounding box.

The corner prediction head is designed based on keypoint detection. Inspired by CornerNet [38], the corner prediction head adopts the CNN to learn the heat map that includes the paired key points information of the object bounding box and then utilizes the Soft-argmax function to calculate the corner coordinates to obtain the object bounding box. Two convolution layers with the same structure are used to obtain the heat map containing the two corners' information of the object. Each convolutional layer includes the structure of four stacked Conv-BN-ReLU layers. Then, the Soft-argmax function is used

to process the heat map to make the heat map can describe the corners' position accurately. Specifically, the function first normalizes the heat map by the Softmax function and then calculates the expected value. The resulting normalized heat map can be viewed as a probability map of the corner at position (x, y). The expected value of the corner position is followed as

$$E = (\sum_{a=1}^{W}\sum_{b=1}^{H} am_{b,a}, \sum_{b=1}^{H}\sum_{a=1}^{W} bm_{b,a}), \tag{9}$$

where m is the normalized heatmap with size $W \times H$ and $E = (e_x, e_y)$ is the corner position.

The corner prediction head encodes the object bounding box estimation into the normalized heat map distribution by retraining the natural spatial structure of the feature map, which can avoid encoding the spatial information into the channel to minimize the loss of spatial information. Therefore, using the corner prediction head in SOM is beneficial to improve the object bounding box's accuracy.

2.3.3. Auxiliary Mask Prediction Head

Given the beneficial performance of mask prediction for improving tracking performance in some tracking tasks [36,39], adding the additional detailed information of the object shape to the SOM facilitates accurate estimation of the object bounding box. Therefore, SOM adds an auxiliary mask prediction head in a position parallel to the corner prediction head, introducing pixel-level supervision into the training to facilitate the optimization module's utilization of more detailed spatial information, further improving the bounding box estimation ability.

The auxiliary mask prediction head needs the strong ability to use spatial detail information. Since the image segmentation task is the pixel-level computer vision task, and U-Net [40] is the most classic algorithm in the segmentation field, the auxiliary mask prediction head is designed based on the U-Net. Specifically, this prediction head is implemented as the U-Net style decoder. First, the feature map containing the template patch and the search region information is upsampled layer by layer. Then, in each layer, the upsampled results are combined with low-level features obtained from the backbone (using stitching and convolution operations) until the feature map has the same resolution as the input image. Finally, the acquired last layer feature map predicts the mask. In particular, to speed up the inference, the mask prediction head is disabled in the test phase to advance the spatial optimization process. More details can be found in the reference [34].

3. Experiments

3.1. Implementation Details

In this work, all experiments were performed using a desktop computer equipped with NVIDIA RTX 3090 GPU and Intel Xeon Silver 4210R CPU. The public hyperspectral dataset provided by Xiong et al. [1] was used for training and testing. The stochastic gradient descent (SGD) method is utilized for training the proposed network. Twenty epochs were trained in total. The learning rate increased linearly from 0.005 to 0.01 in the first 5 epochs and decreased exponentially to 0.0005 in the remaining 15 epochs. We used the multimodality video data composed of RGB and hyperspectral in the training sets as input for the training network. We only adopted the hyperspectral video data of the testing set as the network's input in the testing process. During the testing process, we used the full-band hyperspectral data as the input of the hyperspectral branch in the multimodality fusion information transfer network and the pseudocolor data synthesized by the [1, 8, 16] bands of hyperspectral data as the input of the rest of TMTNet. In addition, we utilized the success plot, the precision plot, the area under the curve (AUC) score of the success rate plot, and the precision rate at the threshold of 20 pixels (DP_20) value of the precision rate plot to evaluate the tracker performance.

3.2. Dataset

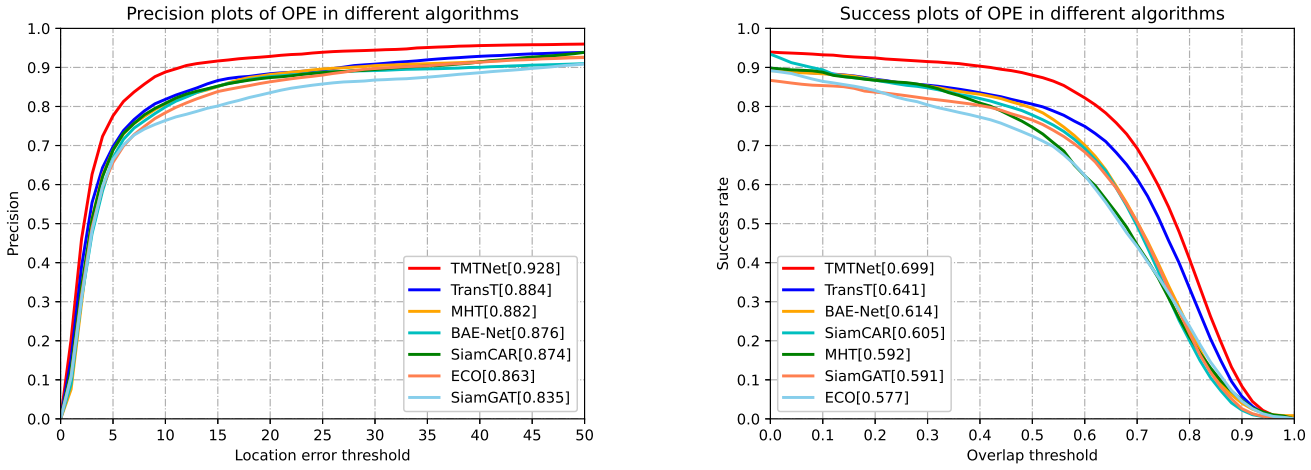
The dataset used in this work is proposed in [1], which contains three types of video data, including hyperspectral video data, false-color video data synthesized from hyperspectral video sequences, and RGB video data taken at the same time from almost the same perspective as hyperspectral video. It is worth noting that to make the RGB sequence and the hyperspectral sequence describe almost the same scene, Xiong et al. [1] carried out a simple coregistration on them. Among them, the labels of hyperspectral and RGB videos are marked separately. In addition, the false-color video data is obtained by converting the hyperspectral video data using the CIE color matching method, which is spatially aligned with the hyperspectral video data, so the label of the false-color video is the same as that of the hyperspectral video. There are eleven challenging factors in the dataset, consisting of low resolution (LR), illumination variation (IV), scale variation (SV), background clutters (BC), occlusion (OCC), motion blur (MB), in-plane rotation (IPR), out-of-plane rotation (OPR), out-of-view (OV), fast motion (FM), and deformation (DEF). The dataset has 40 training set videos and 35 testing set videos in total.

4. Results and Analysis

4.1. Comparison with State-of-the-Art Trackers

In this section, we compare and analyze the performance of the TMTNet tracker with that of the advanced depth color tracker and hyperspectral tracker using the AUC score and the DP_20 value.

Comparison with State-of-the-art Depth Color Trackers. The performance of the TMTNet tracker is compared with that of some advanced color trackers based on deep learning, including TransT [29], SiamCAR [23], SiamGAT [25], and ECO [41], to evaluate the influence of hyperspectral data on tracking performance and the effectiveness of the TMTNet tracker. The TMTNet tracker was run on the hyperspectral video, and the color tracker was run on the false-color video. As shown in Figure 10 and Table 2, the TMTNet tracker's performance is significantly better than that of the compared color tracker and reaches the highest AUC score of 0.699. In addition, Table 3 shows that the TMTNet tracker achieves the best AUC performance compared with the depth color tracker in most challenging scenarios, such as OCC, LR, and BC. In particular, the AUC score of TMTNet is 10.0% higher than that of the best comparative depth color tracker in the BC scenario. It exhibits that hyperspectral data can offer more robust features for the tracking process and also proves that the proposed TMTNet can effectively use hyperspectral data to enhance the ability to cope with challenging scenarios, which indicates the TMTNet's effectiveness.



(**a**) (**b**)

Figure 10. Comparison of success plot and precision plot of tracking results. (**a**) Overall precision plot; (**b**) Overall success plot.

Table 2. AUC score, DP_20 value, and FPS of 7 trackers. The best result is labeled in red.

	TMTNet	TransT	SiamCAR	SiamGAT	ECO	BAE-Net	MHT
AUC	0.699	0.641	0.605	0.591	0.577	0.614	0.592
DP_20	0.928	0.884	0.874	0.835	0.863	0.876	0.882
FPS	12.6	36.7	32.2	20.5	16.8	0.9	1.5

Table 3. The AUC score of 7 trackers in eleven challenging scenarios. The best result is labeled in red.

	TMTNet	TransT	SiamCAR	SiamGAT	ECO	BAE-Net	MHT
BC	0.722	0.622	0.615	0.609	0.606	0.656	0.617
LR	0.619	0.573	0.575	0.487	0.491	0.494	0.481
OCC	0.639	0.628	0.586	0.558	0.537	0.534	0.547
DEF	0.753	0.729	0.642	0.600	0.595	0.681	0.654
IV	0.585	0.558	0.492	0.476	0.547	0.512	0.498
SV	0.677	0.621	0.600	0.587	0.543	0.604	0.570
FM	0.715	0.689	0.704	0.609	0.566	0.612	0.546
IPR	0.782	0.695	0.653	0.643	0.598	0.703	0.643
OPR	0.768	0.700	0.652	0.639	0.601	0.704	0.643
MB	0.672	0.707	0.722	0.604	0.572	0.598	0.565
OV	0.675	0.707	0.696	0.636	0.478	0.605	0.427

Comparison with Hyperspectral Trackers. We also compare the performance of TMTNet with some new hyperspectral object trackers to further verify the proposed method's effectiveness. MHT [1] and BAE-Net [2], excellent hyperspectral trackers, are chosen for comparative experiments. It can be observed from Figure 10 and Table 2 that compared with other hyperspectral trackers, the TMTNet tracker obtained the highest AUC score and DP_20 value. In addition, the AUC score of the TMTNet tracker is also higher than that of the HA-Net tracker (68.7%) [15] that won the Hyperspectral Object Tracking Challenge 2020. Besides, Table 3 also shows that the AUC score of the TMTNet tracker outperforms that of the comparative hyperspectral trackers in 11 challenging scenarios. The results show that the proposed TMTNet can better leverage hyperspectral data to provide robust features under these challenges in the tracking process, enhancing the tracking performance. Moreover, TMTNet is also an extension of our previous work TrTSN [22], the champion scheme of the Hyperspectral Object Tracking Competition 2022, and has achieved similar performance to TrTSN, indicating that the hyperspectral object tracking method designed from the perspective of multi-modality information transfer is flexible, simple, and effective.

Table 2 also shows the FPS of various trackers. It can be found that the proposed tracker's speed is relatively the fastest among the hyperspectral trackers, which can also prove the superiority of the proposed hyperspectral tracker. In addition, Figure 11 shows the qualitative tracking results of some trackers on the sequences of pedestrian2, student, car3, and fruit, which can intuitively compare the tracking performances. These sequences mainly involve the challenging scenes of OCC, IV, SV, DEF, BC, and LR. The above examples show that the proposed TMTNet provides the most accurate boundary frame, which fully demonstrates the TMTNet tracker can effectively deal with various challenging scenarios, proving its effectiveness in hyperspectral tracking.

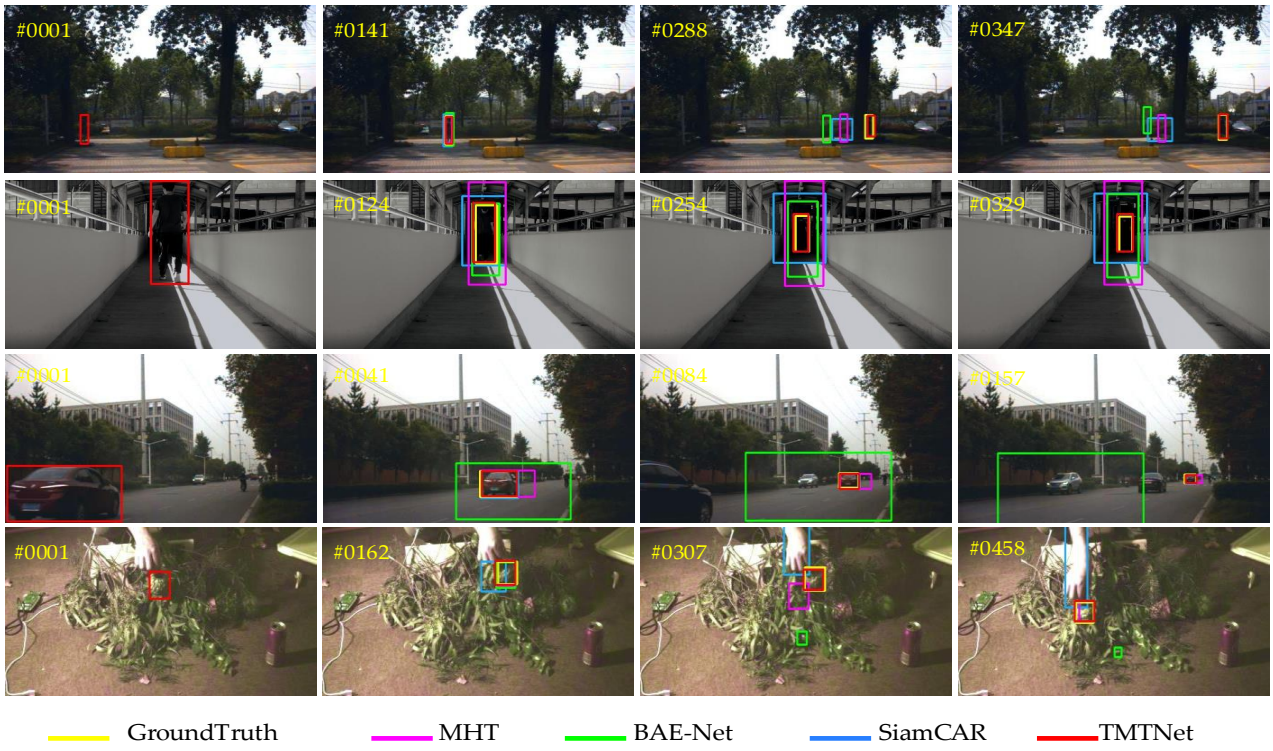


Figure 11. Qualitative result comparison of some trackers on sequences of pedestrian2, student, car3, and fruit.

4.2. Effectiveness of the Transferred Multi-Modality Information

In this work, we propose a Transformer-based multimodality information transfer network (TMTNet) for hyperspectral object tracking, aiming to fully transfer the information of multimodality data composed of RGB data and hyperspectral data to enhance the hyperspectral tracking performance. The transferred multimodality information includes the fusion information of multimodality data composed of RGB and hyperspectral and the RGB modality information. The multimodality fusion information is transferred by the multimodality fusion information transfer subnetwork, which can obtain multimodality-similar fusion information of hyperspectral data to improve tracking performance. The RGB modality information is transferred by the RGB modality information transfer subnetwork, which is used to get robust visual-similar features of hyperspectral data to improve the network's ability to predict the object location in unknown complex scenes. Then, the transferred multimodality fusion information and the RGB modality information are used to predict the object's position jointly.

To prove that the network performance of transferring the multimodality information consisting of the multimodality fusion information and the RGB modality information (achieved by two subnetworks) is better than that of transferring the multimodality fusion information or RGB modality information (using only one subnetwork), we design two TMTNet models without the multimodality fusion information transfer subnetwork or the RGB modality information transfer subnetwork and compare their performance with that of the TMTNet model with two subnetworks (TMTNet). Among them, the TMTNet model that lacks the multimodality fusion information transfer subnetwork but contains the RGB modality information transfer subnetwork is termed as TMTNet_RGB, and the other TMTNet model that does not include the RGB modality information transfer subnetwork but has the multimodality fusion information transfer subnetwork is called TMTNet_fusion.

The experimental results are listed in Table 4. It can be found that the AUC score of the TMTNet tracker (69.9%) is higher than that of the TMTNet_RGB tracker (68.0%) by 1.9%,

and the DP_20 value of the TMTNet tracker (92.8%) is more than that of the TMTNet_RGB tracker (88.7%) by 4.1%. It also can be seen that the AUC score and the DP_20 value of the TMTNet tracker outperform these of the TMTNet_fusion tracker. The above results show that using the transferred multimodality information composed of the multimodality fusion information and the RGB modality information (achieved by two subnetworks) to predict the object's position jointly is conducive to the improvement of the performance of hyperspectral tracking, indicating that the transferred multimodality information in the hyperspectral object tracking is effective.

Table 4. The AUC score and DP_20 value of the TMTNet tracker, TMTNet_RGB tracker, and TMTNet_fusion tracker.

	AUC	**Δ(AUC)**	**DP_20**	**Δ(DP_20)**
TMTNet_RGB	68.0%	-	88.7%	-
TMTNet	69.9%	↑ 1.9%	92.8%	↑ 4.1%
TMTNet_fusion	66.2%	-	88.5%	-
TMTNet	69.9%	↑ 3.7%	92.8%	↑ 4.3%

4.3. Effectiveness of the Transformer-Based Information Interaction Module

Fully fusing different modality information is the key to effectively using the transferred multimodality fusion information to improve the hyperspectral tracking performance. To achieve the multi-modality information fusion, we design an information interaction module based on Transformer (TIIM) in the multimodality fusion information transfer subnetwork to combine the semantic features obtained from Siamese 3D CNN and Siamese 2D CNN branches, which can utilize the Transformer's self-attention mechanism to adaptively obtain the relationship between different modality data for fusing mutimodality information.

To further verify the effectiveness of TIIM, we use the concatenation-based fusion method proposed by Zhu et al. [42] and the cross-based fusion method proposed by Zhang et al. [43] to replace the TIIM in the multimodality fusion information transfer subnetwork respectively and test their performance. The concatenation-based fusion method combines multimodality information by concatenating different modality features, denoted as TMTNet_concat. The cross-based fusion method gets more compact feature representations of multimodality by interactively connecting the depth features from different modalities, termed TMTNet_cross.

In Table 5, the AUC score of the TMTNet tracker (69.9%) outperforms that of the TMTNet_concat tracker (67.6%) and the TMTNet_cross tracker (68.3%) after using the TIIM, while the DP_20 value of the TMTNet tracker (92.8%) is more than that of the TMTNet_concat tracker (88.9%) and the TMTNet_cross tracker (89.5%) by 3.9% and 3.3%, respectively. Experimental results show that the proposed TIIM can effectively fusion different modality information.

Table 5. The AUC score and DP_20 value of the TMTNet tracker, TMTNet_concat tracker, and TMTNet_cross tracker.

	AUC	**Δ(AUC)**	**DP_20**	**Δ(DP_20)**
TMTNet_concat	67.6%	-	88.9%	-
TMTNet	69.9%	↑ 2.3%	92.8%	↑ 3.9%
TMTNet_cross	68.3%	-	89.5%	-
TMTNet	69.9%	↑ 1.6%	92.8%	↑ 3.3%

4.4. Effectiveness of the Response-Level Fusion Method

In the hyperspectral tracking process, selecting an appropriate method to use the multimodality-similar fusion information and visual-similar information obtained from

hyperspectral data to predict the object location jointly is important for effectively utilizing the transferred multimodality information to improve the tracking performance. In this work, we adopt the response-level fusion method to integrate the two sets of response maps obtained by the multimodality-similar fusion information and the visual-similar information into a set of average response maps to predict the object position by using the transferred multimodality information jointly.

To prove the effectiveness of the response-level fusion method, we use the decision-level fusion method, which needs to directly average the final prediction results of the two subnetworks to replace the response-level fusion method in TMTNet to combine the multimodality-similar fusion information and visual-similar information and compare its performance with that of using the response-level fusion method. Among them, the TMTNet model with the decision-level fusion method is termed TMTNet_dec, and the TMTNet model with the response-level fusion method is termed TMTNet_res, which is the actual TMTNet model. The performance of the TMTNet model with different fusion methods is shown in Table 6.

Table 6. The AUC score and DP_20 value of the TMTNet_res tracker and TMTNet_dec tracker.

	AUC	Δ(AUC)	DP_20	Δ(DP_20)
TMTNet_dec	68.0%	-	90.7%	-
TMTNet_res	69.9%	↑1.9%	92.8%	↑ 2.1%

It is evident that the AUC score of the TMTNet_res tracker (69.9%) is over than that of the TMTNet_dec tracker (68.0%) by 1.9%, and the DP_20 value of the TMTNet_res tracker (92.8%) is higher than that of the TMTNet_dec tracker (90.7%) by 2.1%. Experimental results show that using the response-level fusion method in TMTNet to combine the transferred multimodality fusion information and the RGB modality information can effectively improve the tracking network's performance.

4.5. Ablation Study

In this work, the proposed multimodality information transfer network for hyperspectral object tracking mainly includes the subject network and the spatial optimization module, which are adopted to transfer multimodality information and optimize object boundary estimation. There are two subnetworks in the subject network, including the multimodality fusion information transfer subnetwork and the RGB modality information transfer subnetwork, which are used to obtain multimodality-similar fusion information and visual-similar information from hyperspectral data, respectively, and then use the information mentioned above to predict the object location jointly. In this section, we validate the impact of each critical component of TMTNet on final performance. Among them, the multimodality fusion information transfer sub-network is labeled as $MFIT$, the RGB modality information transfer subnetwork is labeled as $RMIT$, and the spatial optimization module is marked as SOM. The ablation study results are listed in Table 7. The model contains $MFIT$, $RMIT$, and SOM in Table 7 is the complete TMTNet model.

Table 7. The AUC score, DP_20 value, FLOPs, Params, and FPS about the ablation study of each critical component in TMTNet.

	AUC	Δ(AUC)	DP_20	Δ(DP_20)	FLOPs (G)	Params (M)	FPS
MFIT	65.4%	-	89.2%	-	1595.2	235.2	16.9
MFIT+RMIT	67.7%	↑ 2.3%	92.7%	↑ 3.5%	1843.0	292.6	13.7
MFIT + RMIT + SOM	69.9%	↑ 4.5%	92.8%	↑ 3.6%	1846.6	333.4	12.6

The symbol *MFIT* represents the multimodality fusion information transfer subnetwork, *RMIT* denotes the RGB modality information transfer subnetwork, and *SOM* is the spatial optimization module.

It can be seen that the TMTNet model with *MFIT* and *RMIT* that adds the RGB modality information based on the transferred mutimodality fusion information, is 2.3% higher than the TMTNet only with *MFIT*, which only transfers the multi-modality fusion information in terms of the AUC score and 3.5% higher in terms of the DP_20 value. The AUC score of the TMTNet model that is adding *SOM* to the TMTNet model with *MFIT* and *RMIT* (69.9%) outperforms the AUC score of the TMTNet model with *MFIT* and *RMIT* (67.7%) by (2.2%), and the DP_20 value of the TMTNet model with *MFIT*, *RMIT* and *SOM* (92.8%) is more than the DP_20 value of the TMTNet model with *MFIT* and *RMIT* (92.7%) by (0.1%).

The results show that the proposed TMTNet model with the multimodality fusion information transfer subnetwork, the RGB information transfer subnetwork, and the spatial optimization module can effectively transfer the multimodality information in the hyperspectral tracking task and optimize object boundary estimation, indicating the designed critical components in the TMTNet model are useful for achieving the performance of the hyperspectral tracking improvement. Although adding components to the tracking model increases the computational complexity of the model and reduces the FPS, it is worth sacrificing a certain amount of calculation and running speed to achieve the model's accuracy improvement in the preliminary exploration stage of hyperspectral object tracking. In the future, we will further explore hyperspectral tracking methods that reduce the model's computational complexity while improving the algorithm's accuracy performance, thus promoting the vigorous development of hyperspectral object tracking.

5. Conclusions

We propose a Transformer-based modality information transfer network for hyperspectral object tracking in this paper, termed as TMTNet, aiming to achieve tracking performance improvement by efficiently transferring the information of multimodality data composed of RGB and hyperspectral. Within this network, two Siamese subnetworks are constructed to transfer the multi-modality fusion information and the robust RGB visual information in the hyperspectral tracking process, respectively, which can improve the ability to predict the object's position accurately by obtaining the multimodality-similar fusion information and the robust visual-similar information from hyperspectral data. Specifically, a Transformer-based information interaction module is designed in the multimodality fusion information transfer subnetwork to fuse multimodality information adaptively by using the Transformer's self-attention mechanism. In addition, a spatial optimization module is added to TMTNet, which further optimizes the object position by fully retaining and utilizing detailed spatial information. The comparison of experimental results with some advanced trackers on the only available hyperspectral benchmark dataset demonstrates the effectiveness of the proposed method.

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

Funding: This work was supported by the National Natural Science Foundation of China (No. 62271159, No. 62071136, No. 62002083, No. 61971153); Heilongjiang Outstanding Youth Foundation (YQ2022F002); Heilongjiang Postdoctoral Foundation (LBH-Q20085 and LBH-Z20051); Fundamental Research Funds for the Central Universities Grant (3072022QBZ0805, 3072021CFT0801 and 3072022CF0808).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset composed of hyperspectral video data, false-color video data, and RGB video data is obtained from https://www.hsitracking.com/ (accessed on 5 April 2021) in this work.

Conflicts of Interest: The authors declare no conflict 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. 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. [CrossRef]
3. 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. [CrossRef]
4. Yang, X.; Wang, Y.; Wang, N.; Gao, X. An Enhanced SiamMask Network for Coastal Ship Tracking. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5612011. [CrossRef]
5. Thomas, M.; Kambhamettu, C.; Geiger, C.A. Motion Tracking of Discontinuous Sea Ice. *IEEE Trans. Geosci. Remote Sens.* **2011**, *49*, 5064–5079. [CrossRef]
6. Fan, H.; Ling, H. Siamese Cascaded Region Proposal Networks for Real-Time Visual Tracking. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7944–7953. [CrossRef]
7. Tochon, G.; Chanussot, J.; Dalla Mura, M.; Bertozzi, A.L. Object Tracking by Hierarchical Decomposition of Hyperspectral Video Sequences: Application to Chemical Gas Plume Tracking. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 4567–4585. [CrossRef]
8. Qian, K.; Zhou, J.; Xiong, F.; Zhou, H.; Du, J. Object Tracking in Hyperspectral Videos with Convolutional Features and Kernelized Correlation Filter. *arXiv* **2018**, arXiv:1810.11819.
9. Nguyen, H.V.; 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. [CrossRef]
10. Uzkent, B.; Hoffman, M.J.; Vodacek, A. Real-Time Vehicle Tracking in Aerial Video Using Hyperspectral Features. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Las Vegas, NV, USA, 26 June–1 July 2016; pp. 1443–1451. [CrossRef]
11. Fang, H.; Liao, Z.; Wang, X.; Chang, Y.; Yan, L. Differentiated Attention Guided Network Over Hierarchical and Aggregated Features for Intelligent UAV Surveillance. *IEEE Trans. Ind. Inform.* **2023** , 1–12. [CrossRef]
12. Fang, H.; Ding, L.; Wang, L.; Chang, Y.; Yan, L.; Han, J. Infrared Small UAV Target Detection Based on Depthwise Separable Residual Dense Network and Multiscale Feature Fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–20. [CrossRef]
13. 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]
14. 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. [CrossRef]
15. Liu, Z.; Wang, X.; Shu, M.; Li, G.; Sun, C.; Liu, Z.; Zhong, Y. An Anchor-Free Siamese Target Tracking Network for Hyperspectral Video. 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. [CrossRef]
16. Xu, N.; Xiao, G.; Zhang, X.; Bavirisetti, D.P. Relative object tracking algorithm based on convolutional neural network for visible and infrared video sequences. In Proceedings of the ICVR 2018: 2018 4th International Conference on Virtual Reality, Hong Kong, China, 24–26 February 2018; pp. 44–49. [CrossRef]
17. Zhang, X.; Ye, P.; Qiao, D.; Zhao, J.; Peng, S.; Xiao, G. Object Fusion Tracking Based on Visible and Infrared Images Using Fully Convolutional Siamese Networks. In Proceedings of the 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019; pp. 1–8.
18. Zhang, X.; Ye, P.; Peng, S.; Liu, J.; Xiao, G. DSiamMFT: An RGB-T fusion tracking method via dynamic Siamese networks using multi-layer feature fusion. *Signal Process. Image Commun.* **2020**, *84*, 115756. [CrossRef]
19. Satar, B.; Hongyuan, Z.; Bresson, X.; Lim, J.H. Semantic Role Aware Correlation Transformer For Text To Video Retrieval. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 1334–1338. [CrossRef]
20. Zhu, Y.; Wang, S.; Huang, Z.; Chen, K. Text Recognition in Images Based on Transformer with Hierarchical Attention. In Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, 22–25 September 2019; pp. 1945–1949. [CrossRef]
21. Le, T.; Nguyen, H.T.; Nguyen, M.L. Vision And Text Transformer For Predicting Answerability On Visual Question Answering. In Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP), Anchorage, AK, USA, 19–22 September 2021; pp. 934–938. [CrossRef]

22. Su, N.; Liu, H.; Zhao, C.; Yan, Y.; Wang, J.; He, J. A Transformer-Based Three-Branch Siamese Network For Hyperspectral Object Tracking. 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. [CrossRef]
23. Guo, D.; Wang, J.; Cui, Y.; Wang, Z.; Chen, S. SiamCAR: Siamese Fully Convolutional Classification and Regression for Visual Tracking. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6268–6276.
24. Chen, Z.; Zhong, B.; Li, G.; Zhang, S.; Ji, R. Siamese Box Adaptive Network for Visual Tracking. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 6667–6676. [CrossRef]
25. Guo, D.; Shao, Y.; Cui, Y.; Wang, Z.; Zhang, L.; Shen, C. Graph Attention Tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9543–9552.
26. Fang, H.; Wang, X.; Liao, Z.; Chang, Y.; Yan, L. A Real-time Anti-distractor Infrared UAV Tracker with Channel Feature Refinement Module. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 1240–1248.
27. He, M.; Li, B.; Chen, H. Multi-scale 3D deep convolutional neural network for hyperspectral image classification. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3904–3908. [CrossRef]
28. 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 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 4277–4286. [CrossRef]
29. Chen, X.; Yan, B.; Zhu, J.; Wang, D.; Yang, X.; Lu, H. Transformer Tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 8122–8131. [CrossRef]
30. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
31. Lanchantin, J.; Wang, T.; Ordonez, V.; Qi, Y. General Multi-label Image Classification with Transformers. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 16473–16483. [CrossRef]
32. Li, M.; Liu, J.; Zheng, C.; Huang, X.; Zhang, Z. Exploiting Multi-view Part-wise Correlation via an Efficient Transformer for Vehicle Re-Identification. *IEEE Trans. Multimedia* **2021** . [CrossRef]
33. Zou, C.; Wang, B.; Hu, Y.; Liu, J.; Wu, Q.; Zhao, Y.; Li, B.; Zhang, C.; Zhang, C.; Wei, Y.; et al. End-to-End Human Object Interaction Detection with HOI Transformer. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11820–11829. [CrossRef]
34. Yan, B.; Zhang, X.; Wang, D.; Lu, H.; Yang, X. Alpha-Refine: Boosting Tracking Performance by Precise Bounding Box Estimation. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5285–5294. [CrossRef]
35. Li, B.; Yan, J.; Wu, W.; Zhu, Z.; Hu, X. High Performance Visual Tracking with Siamese Region Proposal Network. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8971–8980. [CrossRef]
36. Wang, Q.; Zhang, L.; Bertinetto, L.; Hu, W.; Torr, P.H. Fast Online Object Tracking and Segmentation: A Unifying Approach. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1328–1338. [CrossRef]
37. Wang, Z.; Xu, J.; Liu, L.; Zhu, F.; Shao, L. RANet: Ranking Attention Network for Fast Video Object Segmentation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3977–3986. [CrossRef]
38. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *Int. J. Comput. Vis.* **2020**, *128*, 642–656. [CrossRef]
39. Lukežič, A.; Matas, J.; Kristan, M. D3S—A Discriminative Single Shot Segmentation Tracker. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 7131–7140. [CrossRef]
40. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; pp. 234–241.
41. Danelljan, M.; Bhat, G.; Khan, F.S.; Felsberg, M. ECO: Efficient Convolution Operators for Tracking. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6931–6939. [CrossRef]
42. Zhu, Y.; Li, C.; Luo, B.; Tang, J.; Wang, X. Dense Feature Aggregation and Pruning for RGBT Tracking. In Proceedings of the Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 465–472.

43. Zhang, L.; Danelljan, M.; Gonzalez-Garcia, A.; van de Weijer, J.; Shahbaz Khan, F. Multi-Modal Fusion for End-to-End RGB-T Tracking. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 2252–2261. [CrossRef]