remote sensing

MDPI

Article

Hyperspectral Object Detection Based on Spatial–Spectral Fusion and Visual Mamba

Wenjun Li [1,†], Fuqiang Yuan [1,†], Hongkun Zhang [2,*], Zhiwen Lv [3] and Beiqi Wu [1]

[1] College of Transportation, Jilin University, Changchun 130015, China; liwj@jlu.edu.cn (W.L.); yuanfq24@mails.jlu.edu.cn (F.Y.); wubq24@mails.jlu.edu.cn (B.W.)
[2] State Key Laboratory of Automotive Simulation and Control, Jilin University, Changchun 130022, China
[3] School of Communication Engineering, Jilin University, Changchun 130012, China; lvzw22@mails.jlu.edu.cn
* Correspondence: zhanghk@jlu.edu.cn
† These authors contributed equally to this work.

Abstract: Hyperspectral object-detection algorithms based on deep learning have been receiving increasing attention due to their ability to operate without relying on prior spectral information about the target and their strong real-time inference performance. However, current methods are unable to efficiently extract both spatial and spectral information from hyperspectral image data simultaneously. In this study, an innovative hyperspectral object-detection algorithm is proposed that improves the detection accuracy compared to benchmark algorithms and state-of-the-art hyperspectral object-detection algorithms. Specifically, to achieve the integration of spectral and spatial information, we propose an innovative edge-preserving dimensionality reduction (EPDR) module. This module applies edge-preserving dimensionality reduction, based on spatial texture-weighted fusion, to the raw hyperspectral data, producing hyperspectral data that integrate both spectral and spatial information. Subsequently, to enhance the network's perception of aggregated spatial and spectral data, we integrate a CNN with Visual Mamba to construct a spatial feature enhancement module (SFEM) with linear complexity. The experimental results demonstrate the effectiveness of our method.

Keywords: object detection; computer vision; hyperspectral image; spatial–spectral fusion; feature extraction

Citation: Li, W.; Yuan, F.; Zhang, H.; Lv, Z.; Wu, B. Hyperspectral Object Detection Based on Spatial–Spectral Fusion and Visual Mamba. *Remote Sens.* **2024**, *16*, 4482. https://doi.org/10.3390/rs16234482

Academic Editor: Salah Bourennane

Received: 29 October 2024
Revised: 26 November 2024
Accepted: 26 November 2024
Published: 29 November 2024

1. Introduction

Hyperspectral imaging technology is based on the acquisition of image data across numerous narrow spectral bands. It combines imaging technology with spectroscopy to detect the two-dimensional geometric space and one-dimensional spectral information of a target. Due to this characteristic, the underlying material information in hyperspectral images can be applied to object detectors, helping networks to distinguish between objects and complex backgrounds [1]. Hyperspectral imaging technology has been successfully applied in remote sensing [2,3], agriculture [4–6], environmental protection [7,8], medicine, and other fields.

The imaging principle of traditional hyperspectral cameras relies on spectral separation via gratings or prisms, typically consisting of one or more diffraction gratings, optical paths, and detector arrays. Light passes through an input slit and is collimated onto a diffraction grating, which disperses the spectral components in different directions, and a concave mirror focuses this dispersed light toward the detector array [9]. This type of hyperspectral camera can obtain data from hundreds of spectral bands, but the key characteristic—the imaging speed—is poor and it carries high costs. With the continuous development of technology, snapshot cameras based on chip coating principles have overcome the problems related to the slow imaging speeds and large volumes of traditional cameras, and their prices are also approaching the consumer level [10,11]. This type of camera integrates narrowband filters on the surface of a CMOS, allowing the camera to selectively transmit

light of specific wavelengths, making it suitable for selective spectral detection. These cameras have the same imaging speeds as RGB cameras as only the filter is changed. The real-time imaging capabilities of snapshot hyperspectral cameras make them suitable for mobile deployment, such as on unmanned vehicles or in autonomous driving. As the imaging speeds of hyperspectral cameras increase, new demands are being introduced regarding the real-time performance of detection methods.

Previous hyperspectral classification and detection work has mainly focused on the pixel level, relying on spectral information and simple adjacent pixel correlation information [12]. The pixel-based processing method usually requires a long processing period, which is not conducive to the real-time processing of the system. In traditional RGB camera detection methods, object-detection technology is an effective approach that enables real-time inference. Object-detection technology is a branch of computer vision aimed at identifying and locating objects in images or videos, and it is the foundation for other computer tasks, such as real-time object tracking. With the development of deep learning technology and the improvements in computing devices' performance, object-detection technology has been widely applied in various fields, such as autonomous driving [13], transportation surveillance [14], and robot vision [15]. Currently, there are several relevant studies that combine hyperspectral imaging with target-level object-detection technology.

There have been two previous studies on target-level hyperspectral object-detection algorithms. HOD1 was the first object-detection dataset specifically designed for hyperspectral images [12]. The authors introduced a channel attention mechanism into the convolutional neural network to adjust the weights of different spectral channels in the high-dimensional hyperspectral data, and they directly fed the hyperspectral data into the convolutional network for feature extraction, thereby achieving target-level detection. To extract joint spectral and spatial information, S2ADet first applies PCA dimensionality reduction on the raw hyperspectral data as spectral features [16]. The dimensionality reduction process results in the loss of spatial edge information. To compensate for this loss, the authors performed band selection on the raw hyperspectral images, treating the selected data as spatial features. These two sets of features were then treated as different modalities and fed into separate object-detection networks for recognition, and their detection results were aggregated. Due to the local receptive field characteristics of CNNs and the complexity of the aggregation networks, previous studies have been limited by either the detection accuracy or the complexity of the network models.

To overcome the limitations encountered in the existing research, we propose an innovative hyperspectral object-detection algorithm. The main contributions of this study are as follows:

(1) We first propose an edge-preserving dimensionality reduction (EPDR) method based on spatial texture feature weight fusion to ensure that the main spectral features are extracted during the dimensionality reduction process and the key edge and texture information in the image is also preserved;
(2) We propose a multi-scale spatial-feature-enhancement module (SFEM) based on the fusion of a CNN and Mamba, and the experimental results demonstrate the effectiveness of the proposed module;
(3) We analyze the processing speeds of pixel-level and object-level algorithms, demonstrating the superiority of object-level algorithms in terms of efficiency.

2. Related Works

2.1. Pixel-Level Hyperspectral Object Detection

Due to the imaging characteristics of previous devices, the existing research on hyperspectral object-detection technology has mostly focused on pixel-level spectral information, and the detection results usually do not consider the semantic information between multiple objects. This type of algorithm utilizes the spectral dimensional feature differences of images to distinguish between the spectral characteristics of objects and the backgrounds.

According to the presence or absence of prior information, they can be divided into two types: spectral matching detection and anomaly detection.

Spectral matching detection methods usually require the establishment of a spectral database about the target to be detected in advance, as well as the application of similarity measures between the target and prior spectral information for identification. The SAM algorithm, proposed by Kruse et al. in 1993, uses the vector cosine value between the target spectrum and the prior spectrum to determine the similarity in the spectra [17]. In 2002, J. Settle et al. proposed a spectral recognition method based on constraint energy minimization [18]. This method only requires the spectral information of the target object, without the need to establish a database in advance, effectively expanding its application scope. The main idea is to use a finite impulse response filter to match and filter the target, so that the target signal of interest can pass through under specific constraint conditions. The average input energy of the filter caused by unknown signals, such as background signals, is minimized.

Hyperspectral anomaly detection is a binary classification problem that divides an image into the background and the objects of interest [19]. The key to hyperspectral anomaly detection is to analyze pixels with different characteristics from background pixels and identify them as anomalous targets. The RX anomaly-detection algorithm, proposed by Reed et al. in the 1990s, is the benchmark algorithm in this field [20]. This algorithm is based on a statistical model, assuming that the data follow a Gaussian distribution, and it determines whether the target pixel is an anomalous pixel by calculating the Mahalanobis distance between the target pixel and the background pixel. In 2005, Nasrabadi et al. used kernel functions to map hyperspectral data to a high-dimensional feature space based on the RX algorithm, and they improved the accuracy of the algorithm by introducing the concept of nonlinearity. In order to address the need for the adaptability of algorithms to hyperspectral data, Matteoli et al. (2014) proposed the LRX algorithm, which effectively models background data through local adaptive kernel-density-estimation methods, reducing the impact of noise in hyperspectral data and improving the detection accuracy [21]. Due to the spatial sparsity of anomalous targets, some scholars have also used low-rank matrix factorization methods to expand them. For example, Xu et al. proposed a method based on Low-Rank and Sparse Representation (LRaSR). This method uses a low-rank matrix representation of a background dictionary to model background pixels and employs sparse constraints to capture local spectral features, achieving promising results [22]. Ning et al. proposed the Potential Anomaly and Background Dictionary Construction (PAB-DC) algorithm. This algorithm constructs a background dictionary through joint sparse representation and builds a potential anomaly dictionary by analyzing the prior knowledge of anomalous targets in the scene. The use of these dual dictionaries allows for the more accurate differentiation of background, anomaly, and noise pixels from the raw data [23]. Cheng et al. (2020) proposed the GTVLRR algorithm, which jointly extracts spatial and spectral information, preserving the local geometric structures and spatial relationships [24]. With the development of deep learning, Li et al. proposed the CNND algorithm. They were the first to introduce CNN methods into the field of hyperspectral object detection [25]. Subsequently, various deep learning networks emerged.

2.2. Target-Level Hyperspectral Object Detection

With the development of imaging modalities, snapshot hyperspectral cameras are becoming popular in various fields due to their low prices and capacity for instantaneous imaging. The emergence of snapshot hyperspectral cameras has rendered instantaneous detection possible, requiring object-detection algorithms with fast inference capabilities. In order to overcome the limitations of traditional hyperspectral object detection based on the pixel-level detection time, studies have combined hyperspectral image object detection with traditional visible-light object-detection technology to achieve real-time inference functionalities at the target level. These algorithms are typically divided into two stages: the first stage involves extracting effective spatial and spectral features from the original high-

dimensional hyperspectral data, while the second stage focuses on designing a network structure to extract features from the fused data. Typically, the network structures used in the second stage align with those employed in deep-learning-based detection algorithms for RGB images.

Deep-learning-based object detection has become the mainstream method for real-time object detection [26,27]. It is mainly divided into single-stage object-detection algorithms based on regression and two-stage object-detection methods based on candidate regions [28,29]. The two-stage object-detection algorithms mainly include Fast R-CNN [30] and Faster R-CNN [31]. This type of algorithm involves performing fine-grained processing on candidate regions, usually with high detection accuracy. Due to the need for the two-stage processing of candidate region generation and feature classification, the model structure is complex and the real-time performance is poor. The single-stage detection algorithms include the YOLO series [32], SSD [33], Retina-Net [34], and Center-Net [35]. Directly applying these algorithms to the fused data often yields suboptimal results, as the data fused in the first stage typically lack high spatial resolutions and detailed texture features. Spatial attention mechanisms can enhance the model's ability to extract spatial features. Transformers, with the ability to extract global dependencies, can overcome the local receptive field limitations of CNNs; this is the core idea behind many spatial attention mechanisms.

With the powerful global modeling capabilities of the self-attention mechanism, the Transformer has become the dominant algorithm in the field of natural language processing [36]. The Vision Transformer (ViT) represents the application of the Transformer to the field of computer vision [37]. The ViT divides the image into a series of patches and converts each patch into a vector representation as an input sequence. These vectors are then processed through multiple layers of the Transformer encoder, which includes self-attention mechanisms and feedforward neural network layers. This allows the model to capture contextual dependencies across different positions in the image. Since its introduction, numerous improved algorithms based on the ViT architecture have emerged [38,39]. The ViT has also achieved successful application in the field of hyperspectral imaging. Gao et al. proposed a novel Transformer-based hyperspectral target tracking algorithm, leveraging Transformers to extract spectral information for enhanced tracking performance [40]. Ahmad et al. proposed a novel hyperspectral classification model by integrating a wavelet transform with Transformers [41]. To extract feature maps at different scales while establishing global dependencies, the combination of a CNN with a self-attention mechanism has become a common approach. Gong et al. proposed a multi-scale spectral–spatial convolutional model for hyperspectral image classification, integrating a CNN and Transformer [42]. To fully exploit the information in hyperspectral images, Chen et al. proposed a dual-stream collaborative hyperspectral unmixing network based on the ViT and pyramid CNN [43].

The quadratic complexity of the widely used Softmax attention mechanism poses significant challenges when processing high-resolution images or multi-scale features. Numerous works have attempted to reduce the computational cost by introducing local attention windows [44–46] or sparsity [47–49]. Linear attention offers linear complexity and enables the effective modeling of long sequences [50]. However, previous work has shown that linear attention does not always provide satisfactory results, limiting its applicability.

Mamba is a recently proposed state-space model that achieves efficient sequence modeling with linear complexity. Many researchers have attempted to apply the Mamba model to visual tasks and have achieved promising results [51,52]. To adapt it to hyperspectral data inputs, Wang et al. proposed a novel, locally enhanced Mamba network for hyperspectral image classification [53]. Huang et al. proposed a spectral–spatial Mamba model (SS-Mamba) for hyperspectral image classification based on Mamba [54]. The 2D Selective Scan (SS2D) proposed in VMamba relies on the Selective Scanning Spatial State Sequence Model (S6), originally designed for natural language processing tasks, and it successfully addresses the "direction sensitivity" issue in S6. By introducing the Cross-Scan Module

(CSM), SS2D facilitates the extension of S6 to visual models, making it more adaptable to vision tasks. The CSM employs a four-directional scanning strategy to traverse the spatial domain of the image feature map, allowing each pixel in the feature map to integrate information from all positions in various directions. This generates a global receptive field without linearly increasing the computational complexity. Although the SS2D module can integrate global information from different positions, it lacks the ability to perceive multi-scale features. Therefore, in this work, a CNN combined with Mamba is used to implement a spatial feature enhancement module with linear complexity in order to increase the network's spatial extraction ability for fused data.

2.3. Hyperspectral Feature Fusion

Hyperspectral feature fusion aims to retain the spatial and spectral features to the greatest extent while achieving dimensionality reduction. Hyperspectral band selection focuses on selecting a few significant bands to represent the reduced data. Although it preserves the spatial features well, it leads to spectral information loss due to the omission of some spectral bands. In contrast, hyperspectral feature extraction seeks to retain the spectral information while performing dimensionality reduction by obtaining a mapping from the original high-dimensional features to a lower-dimensional space. Typical methods include principal component analysis (PCA) [55] and independent component analysis (ICA) [56]. However, the fused data often lose some spatial information. To address this limitation, edge-preserving dimensionality reduction techniques have been applied, demonstrating excellent performance in the field of hyperspectral processing.

Edge-preserving filtering is an image-processing technique that smooths textures and noise while retaining important content in the image, particularly edge information. This technique has garnered significant attention in the field of computer vision research. Kang et al. innovatively combined edge-preserving filtering techniques with hyperspectral classification. Their experimental results demonstrate that classification methods based on edge-preserving filtering can significantly improve the classification accuracy within a short timeframe [57]. To reduce the computational complexity and improve the classification accuracy, Kang et al. proposed a feature extraction method based on image fusion and recursive filtering. This approach significantly enhanced the accuracy of a support vector machine (SVM) classifier. Compared to other hyperspectral classification methods, it demonstrated superior performance in terms of both classification accuracy and computational efficiency [58]. To address the challenge of directly applying edge-preserving filtering (EPF), with which it is difficult to achieve the complete spatial representation of objects at different scales, Kang et al. proposed a PCA-based EPF method for hyperspectral image (HSI) classification (PCA-EPFs). The results showed that, through the fusion of multi-parameter filtering kernels, this method outperformed traditional EPF-based feature extraction methods and other widely used spectral–spatial classifiers [59]. To address the issue of over-smoothing in the classification maps produced by traditional edge-preserving feature extraction methods, Duan et al. proposed a hyperspectral image-classification method that integrates multiple edge-preserving operators. The algorithm obtains edge-preserving features by implementing different types of edge-preserving operators (EPOs), specifically applying local edge-preserving filtering and global edge-preserving smoothing to the dimensionality-reduced HSI [60].

2.4. Target-Level Object Detection Based on Pre-Fusion Methods

Previous studies on target-level hyperspectral target detection have primarily focused on processing raw hyperspectral data. For instance, one of the earliest approaches involved directly feeding high-dimensional hyperspectral data into object-detection networks, relying solely on channel attention mechanisms to automatically extract the weight relationships between different channels [12]. This method suffers from limitations in terms of the detection accuracy. Moreover, as the number of channels increases, the training time grows significantly, further reducing the practical efficiency of the approach. In subsequent

studies, researchers improved the form of the input data. They treated the band-selected data as spatial information and the PCA-reduced data as spectral information, framing the detection problem as a dual-modal data fusion task. By employing feature interaction modules within the network, they achieved the feature-level fusion of the spectral and spatial information. While this approach improved the detection performance, it introduced a significant computational burden, limiting its efficiency and scalability in practical applications. Edge-preserving dimensionality reduction has proven to be effective in fusing spectral and spatial information. There is currently no research focusing on pre-fusion strategies. Therefore, this study explores the feasibility of pre-fusion approaches for object-level hyperspectral target detection based on edge-preserving dimensionality reduction. This research provides a new perspective on front-end fusion strategies for hyperspectral data processing and validates their potential in target detection tasks.

3. Method

To enable real-time multi-object detection in hyperspectral images, we modified the baseline YOLO model to accommodate hyperspectral image inputs. The flowchart of the proposed detection method is shown in Figure 1. Here, "80 × 80 small" indicates the detection head that is specifically designed for the detection of small objects; it takes a feature map with a size of 80 × 80 as input. Similarly, "40 × 40 medium" and "20 × 20 large" denote the detection heads for medium and large objects, taking feature maps with sizes of 40 × 40 and 20 × 20, respectively.

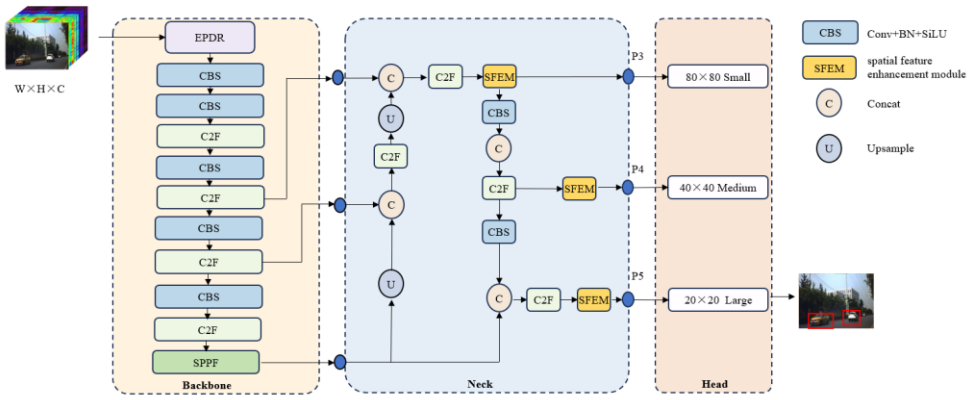

Figure 1. Flow of the proposed detection method.

First, to address the issue of channel redundancy in hyperspectral data, we created a subspace band dimensionality reduction method based on spatial texture-weighted fusion, named EPDR. This module enables the dimensionality reduction of the original hyperspectral data while effectively integrating the spatial and spectral features.

The fused data were processed with the Darknet53 backbone network for feature extraction. In the neck section, we retained the feature pyramid structure, and we incorporated an SFEM module with multi-scale global information-extraction capabilities before each head to enhance the detection accuracy.

3.1. Edge-Preserving Dimensionality Reduction

In hyperspectral data, adjacent bands often exhibit high similarity, while the similarity between non-adjacent bands is generally lower than that among adjacent bands. In Figure 2, the proposed process of spatial–spectral information fusion is shown. The raw hyperspectral data first undergo the ordered band partition step, where the data are evenly partitioned along the spectral dimension. Let $X \in \mathbb{R}^{H \times W \times L}$ denote a hyperspectral image cube, where $X = \{x_1, x_2, x_3, \ldots x_C\}$, and each $x_i \in \mathbb{R}^{N \times L}$ represents a single band of data.

Here, $N = H \times W$ denotes the number of spatial pixels. The process of ordered band partitioning can be expressed in the following formula:

$$G_m \begin{cases} 1, m = 1 \\ \frac{(L - \mathrm{mod}(L,M)) \times m}{M}, 2 \leq \mathrm{m} \leq M - 1 \\ L, m = M \end{cases} \tag{1}$$

where M represents the number of groups, $\mathrm{mod}(\cdot,\cdot)$ denotes the mod operation, and G_m represents the m-th partition point.

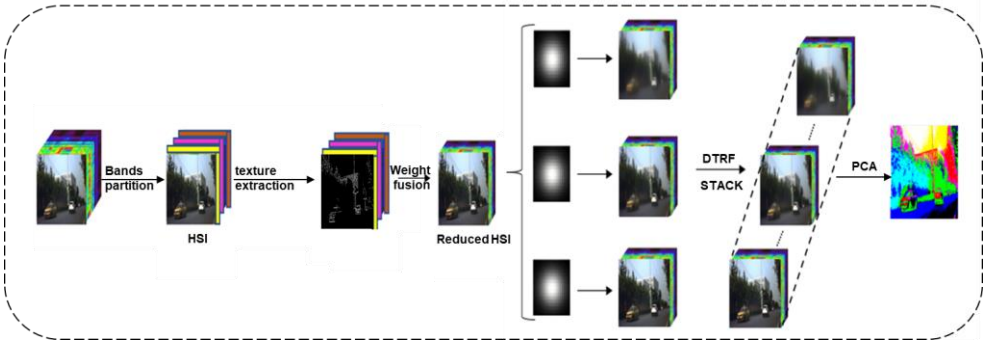

Figure 2. Structure of the improved EPF.

As this partitioning is coarse, it cannot ensure that highly similar bands are accurately grouped together, leading to bands with significant semantic differences being placed in the same group. During the subsequent fusion process, fine details may be obscured, and important features could become blurred. Meanwhile, in order to overcome the problem of inconsistent scales between different bands, we perform weight fusion based on the spatial texture features of different bands, so that bands with better edge features have higher weights during the fusion process.

We use the Sobel method to extract edges from fusion images with bands of different scales. Using *Ei* to denote the image of *Xi* after edge extraction, to evaluate the quality of its texture extraction, we choose the edge continuity as an evaluation indicator. For the edge binary image E, it is assumed that there are m consecutive edge segments, where the i-th segment is composed of a set of pixels as follows:

$$C_i = \left\{ E(x_1^i, y_1^i), E(x_2^i, y_2^i), \cdots E(x_n^i, y_n^i) \right\} \tag{2}$$

The spatial center $(\overline{x}_i, \overline{y}_i)$ of this segment is defined as

$$\begin{cases} \overline{x}_i = \dfrac{1}{n_i} \sum\limits_{k=1}^{n} x_k^i \\ \overline{y}_i = \dfrac{1}{n_i} \sum\limits_{k=1}^{n} y_k^i \end{cases} \tag{3}$$

The distance from each point $E(x_k^i, y_k^i)$ on this segment to the center point $(\overline{x}_i, \overline{y}_i)$ is determined using the following equation:

$$d_k^i = \sqrt{\left(x_k^i - \overline{x}_i\right)^2 + \left(y_k^i - \overline{y}_i\right)^2} \tag{4}$$

The contribution of the pixels on the edge segment to the continuity of the edge segment varies. Pixels closer to the center point have a smaller contribution, while pixels farther away from the center point have a larger contribution. Considering issues, such as the image size and scale, when the distance exceeds a certain value, the contribution of the pixels to the continuity no longer increases with the increase in the distance; rather, it

remains constant. Otherwise, their contribution value will be infinite, leading to numerical calculation errors. Therefore, the contribution of edge segment pixels to the continuity of the edge segment in which they are located is defined as

$$c_k^i = \begin{cases} \frac{d_k^i}{D}, \ d_k^i < D \\ 1, \ d_k^i < D \end{cases} \tag{5}$$

Here, D is the distance threshold, and the size can be selected based on the edge scale required for the application. The sum of the continuity contributions of all pixels on the edge of the segment is C^i

$$C^i = \sum_{k=1}^{n} c_k^i \tag{6}$$

To facilitate the evaluation of the contributions of images from different bands, we normalize C as follows:

$$C_{norm}^i = \frac{C^i - \min(C^i)}{\max(C^i) - \min(C^i)} \tag{7}$$

We calculate the evaluation index S of the image continuity for all connected components using the following equation:

$$S = \frac{\sum_{i=1}^{m} (n_i \times C_{norm}^i)}{n_1} \tag{8}$$

Here, n_1 denotes the number of consecutive components with only one pixel in a single band image. n_i denotes the number of pixels in connected components. S is the texture score calculated for each image; the higher the value of S, the better the texture features of the image.

When fusing the bands in each subspace, we take into account the semantic differences between the features from different bands. We employ a weighted fusion method based on the texture characteristics of images from different spectral bands, ensuring that the features from various bands are effectively integrated for improved fusion accuracy. The formula is as follows:

$$\widetilde{B}_i = \sum_{i=1}^{n} w_i B_i, \ w_i = \frac{e^{S_i}}{\sum_{j=1}^{n} e^{S_j}} \tag{9}$$

Here, B_i denotes the i-th band of the subspace. n is the number of bands contained in the subspace. S_i is the texture feature score of the i-th band image.

Domain transform recursive filtering is a real-time edge-preserving filtering technique that effectively retains edge features during the dimensionality reduction process of hyperspectral images. The filtering process can be divided into two steps: the first step is domain transformation, and the second step is recursive filtering. In the actual filtering process, to control the size and blur of the filter, the domain transformation is often defined as an approximate distance transform. For a given one-dimensional signal I, the domain transformation is defined as

$$U_i = I_0 + \sum_{j=1}^{i} (1 + \frac{\delta_s}{\delta_r} |I_j - I_{j-1}|) \tag{10}$$

where U is the domain-transformed signal; δ_s and δ_r are two parameters used to adjust the smoothness of the filter.

$$J_i = (1 - a^b) I_i + a^b J_{i-1} \tag{11}$$

where J_i is the filter output of the i-th pixel, $a = \exp(-\sqrt{2}/\delta_s) \in [0, 1]$ refers to the feedback coefficient, and b reflects the distance between two neighboring samples U_i and U_{i-1}

in the transform domain. Regarding images, the image is processed by performing the aforementioned 1D operations along each dimension of the image. We refer to the domain transform recursive filtered image as $DTRF(\widetilde{B}, \delta_s, \delta_r)$.

We extract different ambiguities using domain transform recursive filtering with different parameters, and we then stack them to obtain F. We integrate spatial and spectral information by utilizing principal component analysis on the stacked features and retaining the first three principal components:

$$F_k^x = DTRF(\widetilde{B}_k, \delta_s^x, \delta_r^x),\ x = 1, \ldots X,\ k = 1, \ldots, K \tag{12}$$

$$F = \left\{F^1, \ldots, F^X\right\} \tag{13}$$

$$P = PCA(F, 3) \tag{14}$$

Here, δ_s^x and δ_r^x are the xth parameter settings for the domain transform recursive filter. PCA represents the principal component analysis.

3.2. Spatial Feature Enhancement Module (SFEM)

To enhance the original network's ability to extract multi-scale global information, we propose a spatial feature enhancement module (SFEM), which combines the multi-scale characteristics of the CNN with Mamba's linear complexity and global information extraction capabilities. This fusion allows the model to better capture global features at different scales, thereby improving the network's recognition performance. The structure of the SFEM is shown in Figure 3.

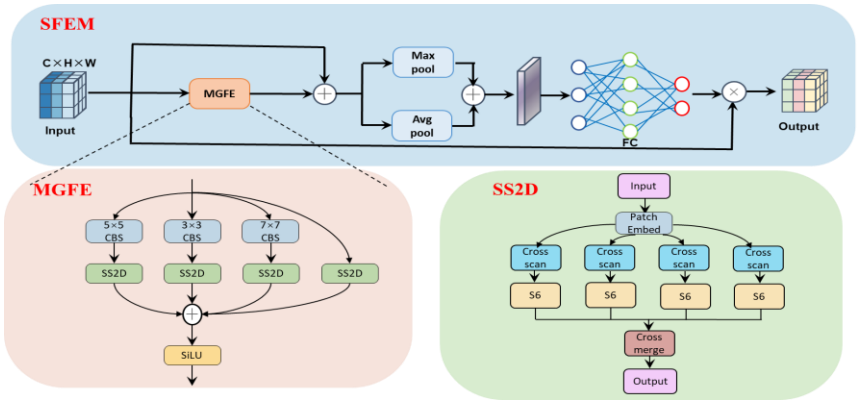

Figure 3. Structure of SFEM.

Firstly, the module uses a multi-granular feature extraction module (MGFE) to extract fine-grained features from the input feature map, using parallel residual blocks with different receptive fields. This design allows the network to capture multi-level information from a fine-grained to macroscopic perspective, enhancing its ability to recognize objects of various sizes and shapes in images. Specifically, the input feature map is first processed through multiple convolutional kernels with different receptive field sizes to extract features at different scales. The output of each convolutional block is fed into the SS2D module, which is specifically designed to analyze and extract the relationships between different positions in the feature map. Through this mechanism, the module identifies and emphasizes the interrelations among key areas within the image, thereby capturing the spatial information that is crucial for understanding the overall scene. This processing approach via the SS2D module not only enhances the granularity of the feature representation but also boosts the network's capability to comprehend the complex interactions among the objects in the scene. In particular, when dealing with partially obscured or overlapping objects, it effectively infers information about the obscured parts, thus improving the accuracy in recognition and analysis. The details of the SS2D module are illustrated in Figure 4.

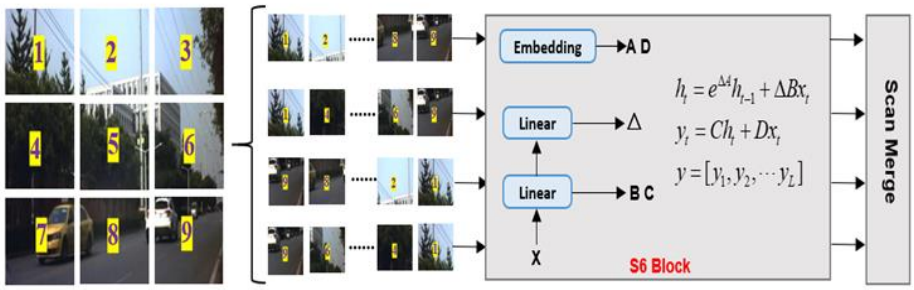

Figure 4. Structure of SS2D.

In order to integrate multi-scale information and enable the network to fully capture the details and structures of images, the system combines features of different scales. The following equations represent this:

$$U_3 = SS2D(\sigma(BN(Conv_{3\times3}(X)))) \tag{15}$$

$$U_5 = SS2D(\sigma(BN(Conv_{5\times5}(X)))) \tag{16}$$

$$U_7 = SS2D(\sigma(BN(Conv_{7\times7}(X)))) \tag{17}$$

$$U = SiLU(U_3 + U_5 + U_7 + X) \tag{18}$$

where $Conv_{3\times3}$, $Conv_{5\times5}$, and $Conv_{7\times7}$ represent convolution kernels of different sizes. BN denotes batch normalization. σ represents the Sigmoid activation function. SS2D represents the 2D Selective Scan module. U represents the output feature map after fusing the multi-scale global information.

In order to effectively compress the feature space and extract more robust and representative feature information, we further process the fused features through max pooling and average pooling operations to obtain feature descriptors. Subsequently, the features are passed through two fully connected layers to limit the model's complexity and aid in generalization. The following equation represents this:

$$\widetilde{M}^s(I) = \sigma(W_2(\delta(W_1(AvgPool(U) \oplus MaxPool(U))))) \tag{19}$$

where δ refers to the RELU function, $W_1 \in \mathbb{R}^{\frac{C}{r}\times C}$, and $W_2 \in \mathbb{R}^{\frac{C}{r}\times C}$.

Through a series of fully connected layers, we generate the final attention map. After undergoing an exponential transformation, this attention map is multiplied by the original input features, effectively highlighting the key features in the image. This method not only enhances the model's sensitivity to important features but also allows for a specific focus on areas that are crucial for interpretation or classification tasks. With such processing, the network is better able to understand and respond to critical information in the image content, thereby achieving higher accuracy and efficiency in various visual tasks. The final attention feature map $\widetilde{E}^s$ is outputted as follows:

$$\widetilde{E}^s = \widetilde{M}^s \otimes I \tag{20}$$

4. Experiments

4.1. Dataset

In our experiments, we used two hyperspectral object-detection benchmark datasets. Detailed information is provided in the following.

4.1.1. HOD1

The HOD1 dataset [12], containing images captured by a push-broom hyperspectral camera, is the first dataset designed for target-level hyperspectral object-detection algorithms. The typical scenario of HOD1 is shown in Figure 5. It covers the visible-to-near-infrared wavelength range, from 400 nm to 1000 nm, and includes 454 hyperspectral

images, each with 96 spectral channels. To verify the specific utility of hyperspectral data, an additional 2048 RGB images with matching spatial resolutions were collected as a control dataset. We maintained the original data-partitioning method, where the train–test ratio was 8:2.

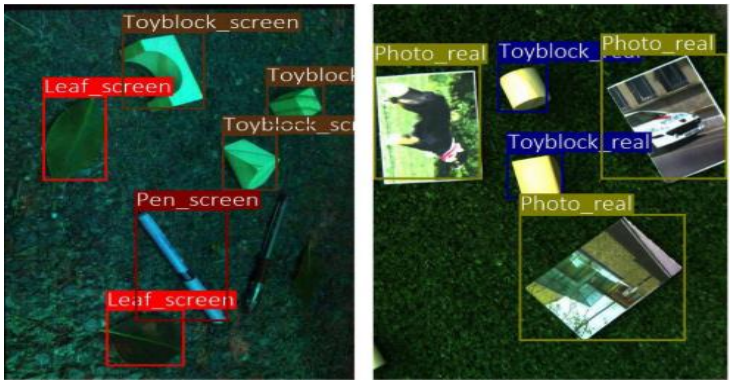

Figure 5. Example from the HOD1 dataset.

These two datasets include various target objects, such as leaves, blocks, pens, and photos. The hyperspectral dataset contains a total of 1657 objects, with an average of 3.64 objects per image. In comparison, the RGB dataset contains a total of 6659 objects, averaging 3.19 objects per image. The similar object-type distribution between the hyperspectral and RGB datasets ensures comparability in experimental studies.

Additionally, in this study, real objects were displayed on an iPad Air and captured simultaneously using hyperspectral and RGB cameras. The results revealed that RGB-only object detection yielded suboptimal outcomes, highlighting the necessity of hyperspectral information for accurate object detection.

4.1.2. HOD3K

The HOD3K [16] dataset originates from the Hyperspectral Object Tracking Challenge, and the images were captured using a snapshot camera. This is the first dataset dedicated to the field of target detection using snapshot hyperspectral cameras. The dataset consists of 3242 annotated images, covering various objects, such as pedestrians, vehicles, and bicycles. These images span 16 spectral channels, ranging from visible to near-infrared wavelengths. We maintained the original dataset-partitioning method, with a ratio of 7:1:2 among the training set, validation set, and test set.

The HOD3K dataset includes multiple scenes, such as campuses, roads, and residential areas. Typical objects and scenes are shown in Figure 6.

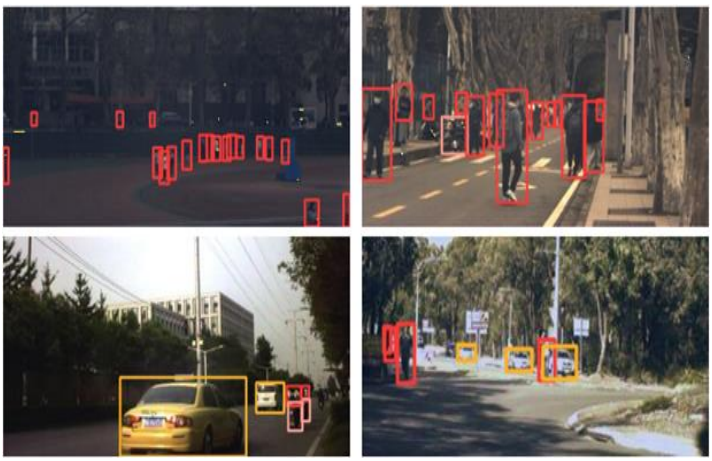

Figure 6. Example from the HOD3K dataset with diverse scenes.

4.2. Experimental Environment

The required environment for the experiment is shown in Table 1. For a fair comparison, all experiments were conducted on a single NVIDIA RTX4090, using the Adam optimizer with a learning rate of 0.01, momentum of 0.973, and a weight decay factor of 0.0005.

Table 1. Test environment parameter configuration.

Platform	Name
CPU	16 vCPU Intel (R) Xeon (R) Platinum 8481C
GPU	RTX 4090D
System	Ubuntu 20.04
Memory	24 GB RAM
GPU acceleration tool	CUDA 11.8

We chose Ubuntu 20.04 as the operating system, with CUDA version 11.8, Python version 1.12.1, and Python 3.9 as the programming environments.

We used Darknet53 as the backbone, with an input image size of 640 × 640. Due to the lack of large-scale datasets suitable for hyperspectral target detection and to test the model's ability to learn independently, we did not use pre-trained weights in this study. All experiments consisted of 50 epochs.

To evaluate the performance of the proposed algorithm, this study employed multiple assessment metrics, including the precision, recall, mean average precision (mAP), and detection speed. The recall is the ratio of the number of true positives to the sum of true positives and false negatives, as defined in Equation (21), where TPs and FNs represent true positives and false negatives, respectively. The accuracy reflects the proportion of true positives to the sum of true positives and false positives; it is detailed in Equation (22), where FP denotes false positives. The mAP is calculated by averaging the precision across all categories, as demonstrated in Equations (23) and (24).

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$Precision = \frac{TP}{TP + FP} \tag{22}$$

$$AP = \int_0^1 P(R) \tag{23}$$

$$mAP = \frac{1}{c} \sum^{c} AP_j \tag{24}$$

4.3. Comparative Experiment

4.3.1. Comparison Between Pixel-Level and Target-Level Detection

To analyze and compare the performance between target-level detection and pixel-level detection, we selected one image from each of the two datasets and performed pixel-level annotation, as shown in Figure 7a,c. It is worth noting that the spatial resolutions of the images in the two datasets were different. The spatial resolution of HOD1 was 859 × 696, while that of HOD3K was 167 × 351. For the pixel-level algorithms, we set the ratio of the training set to the test set to 8:2 and then inputted the new data into the pre-trained network and recorded the testing time. For a fair comparison, all images were reduced to three channels using the PCA algorithm before inference. Here, we only compared the processing times of the algorithms. The comparative algorithms that we selected primarily included CNN3D [61], GSCViT [62], SpeFormer [63], SSFTT [64], MassFormer [65], and GAHT [66].

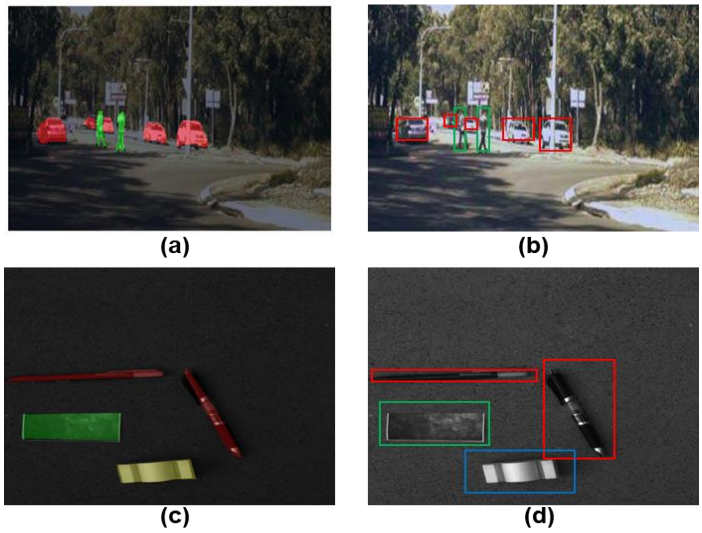

Figure 7. Illustration of different annotation formats; (**a**,**c**) represent pixel-level annotations, while (**b**,**d**) represent object-level annotations.

The detection results of the different algorithms are shown in Figure 8. Based on both datasets, the pixel-level analysis methods had detection times in the range of seconds. Based on the HOD3K dataset, the inference time for the GAHT algorithm was the longest, reaching 5.52 s, while CNN3D had the shortest inference time of 2.13 s. Based on the HOD1 dataset, the inference time for the GAHT algorithm even reached 43.13 s. Regardless of the method, the inference time for the pixel-level analysis methods was significantly larger than that of the object-level detection algorithms based on both datasets. Additionally, it was observed that as the spatial resolution increased, the execution time of the algorithms increased accordingly.

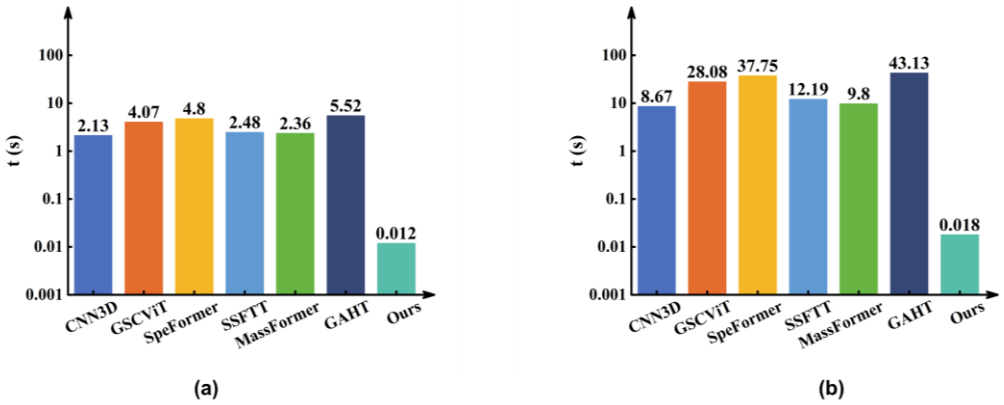

Figure 8. (**a**) Represents the detection time results on the HOD3K dataset, while (**b**) represents the detection time results on the HOD1 dataset.

4.3.2. Comparative Experiments with State-of-the-Art Algorithms

We compared our improved YOLO model with the state-of-the-art object-detection algorithms, including the YOLO series, Faster RCNN, FCOS, RT-DETR, Mamba-YOLO, and S2ADet, which is specifically designed for hyperspectral object detection. The detection results are presented in Tables 2 and 3. By incorporating a spatial feature extraction attention module with linear complexity, our model surpassed the other algorithms in terms of both accuracy and complexity.

Table 2. Comparison of detection results for different methods based on HOD3K. The bold red font represents the highest detection result.

Algorithm	Backbone	Channel	Precision	Recall	mAP_{50}	$mAP_{50:95}$	Params (M)	GFLOPS
RT-DETR	Transformer	3	0.812	0.717	0.794	0.412	61	191.4
Mamba-YOLO [67]	ODMamba	3	0.778	0.689	0.769	0.406	5.98	13.6
FCOS [68]	ResNet50	3	0.723	0.721	0.764	0.397	32.11	161.21
CenterNet [69]	ResNet50	3	0.585	0.41	0.502	0.247	32.6	70.21
RetinaNet [70]	ResNet50	3	0.737	0.475	0.562	0.253	37.96	170
Faster RCNN	ResNet50	3	0.34	**0.845**	0.654	0.31	137	370
YOLOv3 [71]	Darknet53	3	0.799	0.655	0.756	0.405	12.13	19
YOLOv5	Darknet53	3	0.724	0.664	0.786	0.43	2.51	7.2
YOLOv6 [72]	Darknet53	3	0.778	0.683	0.775	0.448	4.23	11.9
YOLOv8	Darknet53	3	0.782	0.675	0.789	0.432	3	8.2
YOLOv9t	Darknet53	3	0.784	0.713	0.766	0.429	2	7.9
YOLOv10n	Darknet53	3	0.733	0.682	0.735	0.409	2.70	8.4
YOLOv11	Darknet53	3	0.742	0.635	0.766	0.446	2.59	6.4
S2ADet [16]	Darknet53	3 + 3	0.739	0.764	0.792	0.438	222.96	169.2
Ours	Darknet53	3	**0.865**	0.722	**0.808**	**0.442**	18	24.2

Table 3. Comparison of detection results for different methods based on HOD1. The bold red font represents the highest detection result.

Algorithm	Backbone	Channel	Precision	Recall	mAP_{50}	$mAP_{50:95}$	Params (M)	GFLOPS
RT-DETR	Transformer	3	0.958	**0.906**	0.948	0.778	61	191.4
Mamba-YOLO	ODMamba	3	0.937	0.844	0.922	0.758	5.98	13.6
FCOS	ResNet50	3	0.942	0.866	0.937	0.776	32.11	161.21
CenterNet	ResNet50	3	0.891	0.848	0.913	0.752	32.6	70.21
RetinaNet	ResNet50	3	0.811	0.749	0.884	0.717	37.96	170
Faster RCNN	ResNet50	3	0.947	0.894	0.947	0.772	137	370
YOLOv3	Darknet53	3	0.925	0.845	0.917	0.744	12.13	19
YOLOv5	Darknet53	3	0.951	0.869	0.942	0.774	2.51	7.2
YOLOv6	Darknet53	3	0.872	0.721	0.891	0.722	4.23	11.9
YOLOv8	Darknet53	3	0.952	0.852	0.944	0.762	3	8.2
YOLOv9t	Darknet53	3	0.944	0.809	0.915	0.759	2	7.9
YOLOv10n	Darknet53	3	0.876	0.792	0.868	0.744	2.70	8.4
YOLOv11	Darknet53	3	0.956	0.866	0.942	0.756	2.59	6.4
S2ADet	Darknet53	3 + 3	0.962	0.872	0.933	0.769	222.96	169.2
Ours	Darknet53	3	**0.964**	0.878	**0.958**	**0.783**	18	24.2

Based on the HOD3K dataset, our proposed method achieved the highest precision, mAP_{50}, and $mAP_{50:95}$. Compared to the baseline YOLOv8, the precision was improved by 8.3%, the recall by 4.7%, the mAP_{50} by 1.9%, and the $mAP_{50:95}$ by 1%, with only a 15 M increase in the computational complexity. By utilizing the Mamba module with linear attention to capture global dependencies, our model could obtain more comprehensive feature information compared to the baseline object-detection network, while the network complexity only increased linearly. Additionally, compared to the state-of-the-art S2ADet network, which was specifically designed for hyperspectral object detection, our method achieved a 1.6% increase in the mAP_{50} and a 12.6% improvement in precision, while using only one-twelfth of the parameters. Meanwhile, the S2ADet network, which incorporates multiple Transformer-based spectral–spatial aggregation modules, exhibited high network complexity. The experimental results fully demonstrate the effectiveness of our proposed method.

Based on the HOD1 dataset, the detection results of the different algorithms were relatively similar. Our proposed method achieved the highest detection accuracy, precision, mAP_{50}, and $mAP_{50:95}$. Compared to the baseline model, the improved model achieved significant enhancements across various metrics, including a 1.2% increase in precision,

a 2.6% increase in recall, a 1.4% improvement in the mAP_{50}, and a 2.1% increase in the $mAP_{50:95}$. Additionally, compared to the state-of-the-art S2ADet network, designed for hyperspectral object detection, our method achieved a 2.5% improvement in the mAP_{50} and a 1.4% improvement in the $mAP_{50:95}$.

4.3.3. Visual Comparison of Detection Results

To provide a more intuitive comparison, we selected several groups of images from the test dataset for an analysis.

As shown in Figure 9, our proposed algorithm achieved comparable detection performance to the state-of-the-art S2ADet. Furthermore, in certain occlusion scenarios, our algorithm outperformed some of the mainstream methods, including YOLOv5, YOLOv8, RT-DETR, and S2ADet. This improvement can be attributed to its effective extraction of global dependencies and the fusion of multi-scale features, which significantly enhance the robustness and accuracy of the detection process in complex conditions.

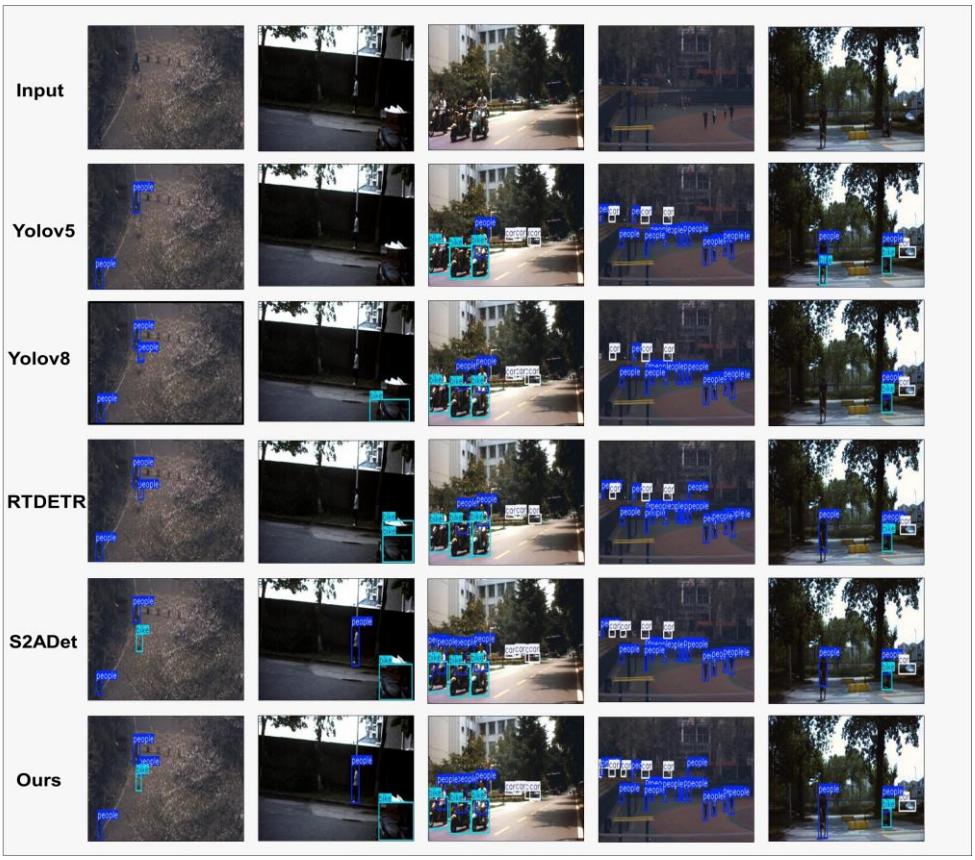

Figure 9. Visualization of detection results for different algorithms. For analysis purposes, we converted the hyperspectral image into a pseudo-color image for display.

4.4. Ablation Experiment

To validate the impact of each module in our proposed method on the detection results, we conducted a comprehensive ablation study on the HOD3K dataset. The baseline network for the ablation study was YOLOv8n, with the training set to 200 epochs. The detection results obtained by sequentially incorporating our modules into YOLOv8n are shown in Table 4. In the table, TL represents transfer learning, which indicates the detection results obtained after incorporating pre-trained weights.

Table 4. Ablation experiment based on different modules. √ indicates "this module is enabled. The bold red font represents the highest detection result.

YOLOv8n	EPDR	SFEM	TL	mAP_{50}	$mAP_{50:95}$	Precision	Recall
√				0.561	0.296	0.741	0.48
√	√			0.781	0.432	0.842	0.711
√		√		0.792	0.512	0.851	0.749
√	√	√		0.808	0.442	0.865	0.722
√	√	√	√	**0.845**	**0.534**	**0.877**	**0.768**

The experimental results demonstrate that the aggregation of spectral and spatial information enhanced the network's ability to extract features, and the use of Mamba's spatial attention mechanism to extract global information at multiple scales could also improve the network's detection capabilities.

4.4.1. Effectiveness of the Proposed EPDR Module

To validate the effectiveness of the EPDR module, we conducted a comprehensive ablation study, seeking to evaluate the impact of various input image-processing methods on the detection performance. The results are shown in Table 5. Considering the network complexity and accuracy, we selected the baseline YOLOv8n as the validation algorithm. First, we directly inputted the original 16-band hyperspectral data into the network as a baseline for evaluation. The increase in the number of channels significantly impacted the inference speed of the network. Additionally, the poor imaging quality in certain bands adversely affected the detection accuracy of the network.

Table 5. The results of different channel fusion methods based on HOD3K. The bold red font represents the highest detection result.

Algorithm	Channel	Time (ms)	People	Bike	Car	mAP_{50}	$mAP_{50:95}$	Precion	Recall
Raw Data	16	9.9	0.386	0.629	0.667	0.561	0.296	0.741	0.48
PCA	3	1.4	0.628	0.652	0.874	0.718	0.414	0.807	0.62
PCA + EPF	3	1.6	0.701	0.711	0.892	0.768	0.429	0.821	0.642
FNGBS	3	3.5	0.638	0.704	0.869	0.737	0.421	0.775	0.688
EFDPC	3	1.8	0.431	0.691	0.876	0.666	0.381	0.743	0.597
ASPS	3	1.8	0.697	0.695	0.863	0.751	0.424	0.713	0.705
Ours	3	1.7	**0.717**	**0.724**	**0.902**	**0.781**	**0.432**	**0.842**	**0.711**

Band selection is currently the most effective method for the manual extraction of hyperspectral band features. Its objective is to select channels that are information-rich and minimally correlated from all available channels. While the dimensionality-reduced images obtained through band selection generally retain good spatial features, they often suffer from a loss of spectral characteristics. To assess the impact of data pre-processing operations, such as band selection, on the real-time performance of target detection, we selected three fast unsupervised hyperspectral band-selection algorithms and tested their execution times. Firstly, we tested the execution speeds of the different methods, and the results are shown in the Table 5. From this, it can be seen that the speed in selecting different numbers of frequency bands using the same method was very similar, but the speed difference between different methods was significant. For example, the execution speeds of algorithms, such as EFDPC, were approximately 0.3 s; this makes it difficult to integrate these algorithms into mobile terminals, such as autonomous vehicles or unmanned aerial vehicles, for real-time target detection. Due to the selection of channels through band selection, the use of band-selection algorithms can also improve the detection accuracy of the network. For example, using ASPS for band selection could improve the model accuracy by 19%.

PCA reduces the dimensionality of hyperspectral data by computing a covariance matrix and extracting the principal components, retaining only the top components that account for the largest variance. When using the PCA dimensionality reduction algorithm to fuse multi-dimensional channels, the model's accuracy was improved by 15.7% compared to the benchmark algorithm. However, dimensionality-reduced data often suffer from the significant loss of spatial information. To overcome this spatial loss, the addition of edge-preserving filtering was considered, which could increase the mAP_{50} by 5%.

By introducing a weighted fusion method based on spatial texture features for different spectral bands in the subspace, the phenomenon of feature fading caused by large semantic differences between images in different spectral bands can be overcome. Compared with the PCA dimensionality reduction method, the accuracy was improved by 6.3% after applying this technique.

4.4.2. Effectiveness of the Proposed SFEM

To validate the effectiveness of the proposed SFEM, we selected several scenes with targets of varying scales, aiming to test its attention allocation maps. Figure 10 illustrates the comparative detection results in different occlusion scenarios when using the baseline model and the baseline model with the SFEM. The first row shows the original images, the second row displays the detection results for the baseline model, and the third row presents the detection results after incorporating the SFEM into the baseline model.

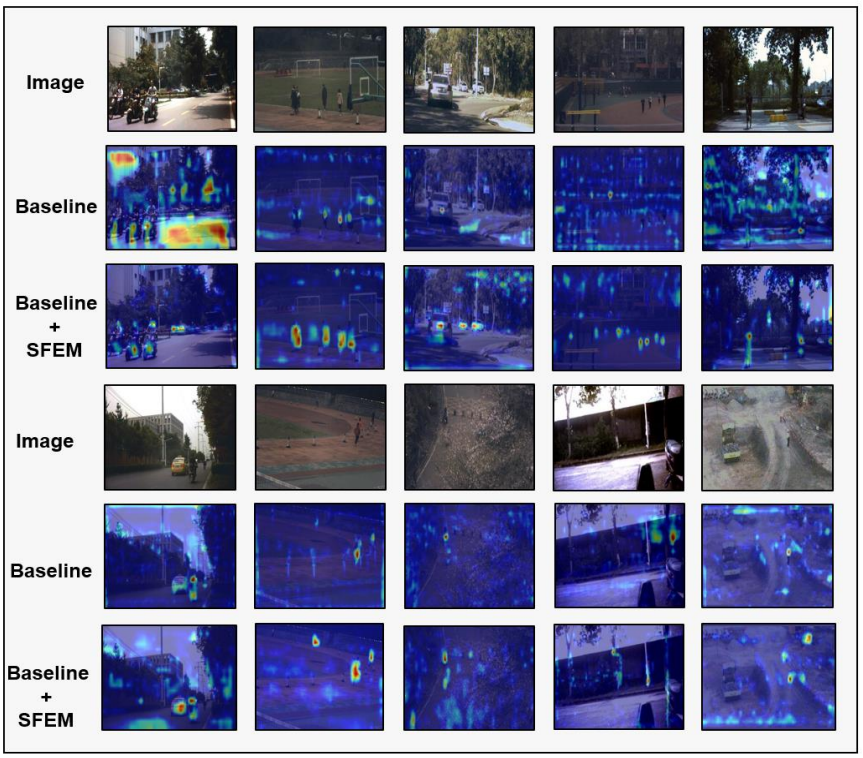

Figure 10. Grad-CAM visualization results. We compare the visualization results of the SFEM-integrated network with those of the baseline (YOLOv8n).

The input to the network consisted of data processed by the EPDR module; for visualization purposes, the detection results were overlaid onto pseudo-color images. The feature maps processed by the EPDR module could lose some spatial texture features, limiting the baseline network's ability to extract such low-spatial-information data and resulting in less focused attention allocation. The SFEM extracts multi-scale global dependencies, enabling the network to capture complex spatial and contextual relationships with different levels of detail. This capability is particularly critical in scenarios involving objects with varying sizes, complex backgrounds, or occlusions, where relying solely on local features

may result in the failure to achieve accurate detection. By integrating global dependencies across multiple scales, the network gains a more comprehensive understanding of the scene, effectively modeling both local and global details, thereby enhancing its ability to represent spatial features. As shown in Figure 7, the improved network demonstrated more focused attention allocation toward the detection objects compared to the baseline network. The integration of multi-scale information in the SFEM enabled the network to dynamically adapt to objects of different scales. For images containing multi-scale targets, such as the third image in the first row and the second image in the second row, the improved network allocated more balanced attention weights to targets of different scales compared to the baseline network. This ensured that small object features were adequately emphasized, even in the presence of larger or more prominent targets in the scene. The extraction of global dependencies allows the network to move beyond local features by leveraging global information to analyze the relationship between the target and its surrounding pixels. In occlusion scenarios, this global perspective significantly enhances the network's detection capabilities, as it can infer and identify partially visible targets using contextual information. By bridging the gap between local details and the global context, the network demonstrates greater robustness and higher detection accuracy when applied to partially occluded or complex background environments. For instance, for the third and fourth images of the second row in Figure 7, the improved model showed a stronger focus on occluded targets compared to the baseline network.

5. Discussion

From the experiments described above, it is evident that the ability to handle high-dimensional raw data is a critical factor influencing the detection accuracy in target-level hyperspectral object-detection algorithms. Without dimensionality reduction, the strong similarity between adjacent hyperspectral bands can lead to the problem of dimensionality. As the number of bands increases, the detection accuracy decreases. Regarding methods that use PCA directly for dimensionality reduction, although they improve the detection accuracy compared to directly inputting the raw data, the reduced data often lose significant amounts of spatial information, resulting in poor spatial texture features, which ultimately limits the detection accuracy. Similar to PCA, band selection is another dimensionality reduction method for hyperspectral data. It primarily focuses on spatial feature extraction. By discarding certain bands directly, this approach results in the loss of spectral features, which ultimately limits its detection accuracy. In S2ADet, to fully leverage the spectral and spatial information, the authors used PCA and band selection to generate spatial and spectral feature maps, respectively, which were then fused using a Transformer module, yielding promising results. However, due to the quadratic complexity of the Transformer, this approach results in a substantial number of parameters. We adopted the concept of edge-preserving filtering for dimensionality reduction, which allowed for the significant retention of spatial features during PCA-based dimensionality reduction. Additionally, to avoid potential feature fading caused by simple averaging between the subspace bands, we proposed an evaluation method for spatial texture features and performed weighted fusion based on the texture characteristics of different spectral bands. The experimental results demonstrate the effectiveness of our approach.

The experimental results indicate that, for deep learning networks, two-stage algorithms generally achieve higher recall than one-stage algorithms, albeit at the cost of increased complexity. Due to the use of a self-attention mechanism, RT-DETR achieves favorable results in terms of its detection accuracy and recall. However, its parameter count reaches 61 M, which is several to tens of times larger than that of typical CNN networks. Regarding algorithms based on the Mamba architecture, the lack of multi-scale perception limits their detection performance. We propose a method that combines a CNN with Vision Mamba, integrating the CNN's multi-scale perceptual capabilities with Mamba's global dependency extraction abilities. By applying a selective attention mechanism to spatial

information, this approach significantly improves the detection accuracy while maintaining linear computational complexity.

The current research has certain limitations. First, the existing dataset was collected using a general-purpose snapshot hyperspectral camera, with a limited number of bands that had not been specifically optimized for feature selection. This resulted in suboptimal bands to distinguish target objects, thereby affecting the detection accuracy. Secondly, at the algorithm level, the current methods perform data preprocessing and network analysis as separate stages, which poses limitations for practical deployment. In the future, end-to-end algorithms will be needed to address this limitation and improve the efficiency and feasibility in real-world applications. Moreover, this study only validated the applicability of pre-fusion for object detection, lacking a comprehensive investigation into other data fusion methods. Future work should explore and evaluate the effectiveness of alternative fusion strategies to further enhance the detection performance.

6. Conclusions

In this study, an innovative target-level hyperspectral object-detection method is proposed and, for the first time, the feasibility of pre-fusion is validated. We utilized edge-preserving dimensionality reduction to achieve the aggregation of spectral and spatial information at the front end, effectively overcoming the high complexity of feature-level fusion. The effectiveness of our fusion method provides a new perspective for future research on hyperspectral algorithms. When integrating spatial and spectral data, we applied a weighted fusion method based on image texture features to improve the subspace band fusion process, thereby overcoming the feature degradation caused by average fusion. To address the challenge of poor spatial characteristics in the fused data, we designed a multi-scale spatial enhancement module combining a CNN and Mamba. This module introduces a linear-complexity global-information-extraction mechanism, significantly improving the modeling capabilities for spatial features. Specifically, the multi-scale spatial enhancement module leverages a CNN to capture multi-scale features, while efficiently extracting global spatial dependencies through Mamba, thereby preserving the spatial information during the feature fusion process. Furthermore, our study demonstrates the significant advantage of object-level detection techniques over pixel-level methods in terms of real-time performance. The current algorithms still adopt a two-stage execution approach. Future research should focus on designing more efficient end-to-end detection frameworks while further optimizing the computational efficiency and detection accuracy of the algorithms. With the continuous advancement of snapshot hyperspectral camera technology, the proposed framework will provide a valuable reference and directions for future algorithmic improvements.

Author Contributions: Methodology, F.Y. and B.W.; Software, F.Y. and Z.L.; Validation, W.L. and H.Z.; Data curation, B.W.; Writing—original draft, F.Y.; Writing—review & editing, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. 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]
2. Ömrüuzun, F.; Çetin, Y.Y.; Leloğlu, U.M.; Demir, B. A Novel Semantic Content-Based Retrieval System for Hyperspectral Remote Sensing Imagery. *Remote Sens.* **2024**, *16*, 1462. [CrossRef]
3. Zheng, H.; Li, D.; Zhang, M.; Gong, M.; Qin, A.K.; Liu, T.; Jiang, F. Spectral Knowledge Transfer for Remote Sensing Change Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4501316. [CrossRef]

4. Zhang, W.; Li, Z.; Li, G.; Zhuang, P.; Hou, G.; Zhang, Q.; Li, C. GACNet: Generate Adversarial-Driven Cross-Aware Network for Hyperspectral Wheat Variety Identification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5503314. [CrossRef]
5. Chen, M.; Feng, S.; Zhao, C.; Qu, B.; Su, N.; Li, W.; Tao, R. Fractional Fourier-Based Frequency-Spatial–Spectral Prototype Network for Agricultural Hyperspectral Image Open-Set Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5514014. [CrossRef]
6. Neri, I.; Caponi, S.; Bonacci, F.; Clementi, G.; Cottone, F.; Gammaitoni, L.; Figorilli, S.; Ortenzi, L.; Aisa, S.; Pallottino, F.; et al. Real-Time AI-Assisted Push-Broom Hyperspectral System for Precision Agriculture. *Sensors* **2024**, *24*, 344. [CrossRef]
7. Darvishi, P.; Karimi, D. Environmental Studies of the Khorramrood River in Iran, Based on Transformed High-Resolution Remotely Sensed Spectroscopic Data. *Egypt. J. Remote Sens. Space Sci.* **2024**, *27*, 298–316. [CrossRef]
8. Liu, B.; Li, T. A Machine-Learning-Based Framework for Retrieving Water Quality Parameters in Urban Rivers Using UAV Hyperspectral Images. *Remote Sens.* **2024**, *16*, 905. [CrossRef]
9. Yang, Z.; Albrow-Owen, T.; Cai, W.; Hasan, T. Miniaturization of Optical Spectrometers. *Science* **2021**, *371*, eabe0722. [CrossRef] [PubMed]
10. Geelen, B.; Blanch, C.; Gonzalez, P.; Tack, N.; Lambrechts, A. A Tiny VIS-NIR Snapshot Multispectral Camera. In *Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VIII, Proceedings of the SPIE OPTO 2015, San Francisco, CA, USA, 13 March 2015*; Von Freymann, G., Schoenfeld, W.V., Rumpf, R.C., Helvajian, H., Eds.; SPIE: Bellingham, WA, USA, 2015; p. 937414.
11. Geelen, B.; Tack, N.; Lambrechts, A. A Compact Snapshot Multispectral Imager with a Monolithically Integrated Per-Pixel Filter Mosaic. In *Advanced Fabrication Technologies for Micro/Nano Optics and Photonics VII, Proceedings of the SPIE MOEMS-MEMS 2014, San Francisco, CA, USA, 7 March 2014*; Von Freymann, G., Schoenfeld, W.V., Rumpf, R.C., Eds.; SPIE: Bellingham, WA, USA, 2014; p. 89740L.
12. Yan, L.; Zhao, M.; Wang, X.; Zhang, Y.; Chen, J. Object Detection in Hyperspectral Images. *IEEE Signal Process. Lett.* **2021**, *28*, 508–512. [CrossRef]
13. Ding, N.; Zhang, C.; Eskandarian, A. SalienDet: A Saliency-Based Feature Enhancement Algorithm for Object Detection for Autonomous Driving. *IEEE Trans. Intell. Veh.* **2024**, *9*, 2624–2635. [CrossRef]
14. Shao, Z.; Wang, L.; Wang, Z.; Du, W.; Wu, W. Saliency-Aware Convolution Neural Network for Ship Detection in Surveillance Video. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 781–794. [CrossRef]
15. Fu, J.; Zong, L.; Li, Y.; Li, K.; Yang, B.; Liu, X. Model Adaption Object Detection System for Robot. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; IEEE: New York, NY, USA, 2020; pp. 3659–3664.
16. He, X.; Tang, C.; Liu, X.; Zhang, W.; Sun, K.; Xu, J. Object Detection in Hyperspectral Image via Unified Spectral-Spatial Feature Aggregation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5521213. [CrossRef]
17. Chang, C.-I. An Information-Theoretic Approach to Spectral Variability, Similarity, and Discrimination for Hyperspectral Image Analysis. *IEEE Trans. Inf. Theory* **2000**, *46*, 1927–1932. [CrossRef]
18. Settle, J. On Constrained Energy Minimization and the Partial Unmixing of Multispectral Images. *IEEE Trans. Geosci. Remote Sens.* **2002**, *40*, 718–721. [CrossRef]
19. Su, H.; Wu, Z.; Zhang, H.; Du, Q. Hyperspectral Anomaly Detection: A Survey. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 64–90. [CrossRef]
20. Reed, I.S.; Yu, X. Adaptive Multiple-Band CFAR Detection of an Optical Pattern with Unknown Spectral Distribution. *IEEE Trans. Acoust. Speech Signal Process.* **1990**, *38*, 1760–1770. [CrossRef]
21. Matteoli, S.; Veracini, T.; Diani, M.; Corsini, G. A Locally Adaptive Background Density Estimator: An Evolution for RX-Based Anomaly Detectors. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 323–327. [CrossRef]
22. Xu, Y.; Wu, Z.; Li, J.; Plaza, A.; Wei, Z. Anomaly Detection in Hyperspectral Images Based on Low-Rank and Sparse Representation. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 1990–2000. [CrossRef]
23. Huyan, N.; Zhang, X.; Zhou, H.; Jiao, L. Hyperspectral Anomaly Detection via Background and Potential Anomaly Dictionaries Construction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 2263–2276. [CrossRef]
24. Cheng, T.; Wang, B. Graph and Total Variation Regularized Low-Rank Representation for Hyperspectral Anomaly Detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 391–406. [CrossRef]
25. Li, W.; Wu, G.; Du, Q. Transferred Deep Learning for Anomaly Detection in Hyperspectral Imagery. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 597–601. [CrossRef]
26. Gong, M.; Zhao, H.; Wu, Y.; Tang, Z.; Feng, K.-Y.; Sheng, K. Dual Appearance-Aware Enhancement for Oriented Object Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5602914. [CrossRef]
27. Yao, Y.; Cheng, G.; Lang, C.; Yuan, X.; Xie, X.; Han, J. Hierarchical Mask Prompting and Robust Integrated Regression for Oriented Object Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *10*, 3444795. [CrossRef]
28. Wu, A.; Deng, C. TIB: Detecting Unknown Objects via Two-Stream Information Bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 611–625. [CrossRef] [PubMed]
29. Wu, A.; Deng, C.; Liu, W. Unsupervised Out-of-Distribution Object Detection via PCA-Driven Dynamic Prototype Enhancement. *IEEE Trans. Image Process.* **2024**, *33*, 2431–2446. [CrossRef] [PubMed]
30. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: New York, NY, USA, 2015; pp. 1440–1448.
31. 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]

32. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 779–788.
33. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9905, pp. 21–37.
34. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018, Salt Lake City, UT, USA, 18–23 June 2018.
35. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE: New York, NY, USA, 2019; pp. 6568–6577.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need 2023. Available online: https://arxiv.org/pdf/1706.03762 (accessed on 25 November 2024).
37. 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 ICLR 2021, Vienna, Austria, 4 May 2021.
38. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the ICLR 2021, Vienna, Austria, 4 May 2021.
39. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024.
40. 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*, 5528211. [CrossRef]
41. Ahmad, M.; Ghous, U.; Usama, M.; Mazzara, M. WaveFormer: Spectral–Spatial Wavelet Transformer for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 5502405. [CrossRef]
42. Gong, Z.; Zhou, X.; Yao, W. MultiScale Spectral–Spatial Convolutional Transformer for Hyperspectral Image Classification. *IET Image Process.* **2024**, *18*, 4328–4340. [CrossRef]
43. Chen, J.; Yang, C.; Zhang, L.; Yang, L.; Bian, L.; Luo, Z.; Wang, J. TCCU-Net: Transformer and CNN Collaborative Unmixing Network for Hyperspectral Image. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 8073–8089. [CrossRef]
44. Dong, X.; Bao, J.; Chen, D.; Zhang, W.; Yu, N.; Yuan, L.; Chen, D.; Guo, B. CSWin Transformer: A General Vision Transformer Backbone with Cross-Shaped Windows. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: New York, NY, USA, 2022; pp. 12114–12124.
45. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 9992–10002.
46. Hassani, A.; Walton, S.; Li, J.; Li, S.; Shi, H. Neighborhood Attention Transformer. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE: New York, NY, USA, 2023; pp. 6185–6194.
47. Zhu, L.; Wang, X.; Ke, Z.; Zhang, W.; Lau, R. BiFormer: Vision Transformer with Bi-Level Routing Attention. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; IEEE: New York, NY, USA, 2023; pp. 10323–10333.
48. Xia, Z.; Pan, X.; Song, S.; Li, L.E.; Huang, G. Vision Transformer with Deformable Attention. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; IEEE: New York, NY, USA, 2022; pp. 4784–4793.
49. Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; Shao, L. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; IEEE: New York, NY, USA, 2021; pp. 548–558.
50. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention. In Proceedings of the International Conference on Machine Learning, Online, 13–18 July 2020.
51. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual State Space Model 2024. Available online: https://arxiv.org/abs/2401.10166 (accessed on 26 May 2024).
52. Yang, C.; Chen, Z.; Espinosa, M.; Ericsson, L.; Wang, Z.; Liu, J.; Crowley, E.J. PlainMamba: Improving Non-Hierarchical Mamba in Visual Recognition. *arXiv* **2024**, arXiv:2403.17695.
53. Wang, C.; Huang, J.; Lv, M.; Du, H.; Wu, Y.; Qin, R. A Local Enhanced Mamba Network for Hyperspectral Image Classification. *Int. J. Appl. Earth Obs. Geoinf.* **2024**, *133*, 104092. [CrossRef]
54. Huang, L.; Chen, Y.; He, X. Spectral-Spatial Mamba for Hyperspectral Image Classification. *Remote Sens.* **2024**, *16*, 2449. [CrossRef]
55. Fauvel, M.; Chanussot, J.; Benediktsson, J.A. Kernel Principal Component Analysis for the Classification of Hyperspectral Remote Sensing Data over Urban Areas. *EURASIP J. Adv. Signal Process.* **2009**, *2009*, 783194. [CrossRef]
56. Wang, J.; Chang, C.-I. Independent Component Analysis-Based Dimensionality Reduction with Applications in Hyperspectral Image Analysis. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 1586–1600. [CrossRef]

57. Kang, X.; Li, S.; Benediktsson, J.A. Spectral–Spatial Hyperspectral Image Classification with Edge-Preserving Filtering. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 2666–2677. [CrossRef]
58. Kang, X.; Li, S.; Benediktsson, J.A. Feature Extraction of Hyperspectral Images with Image Fusion and Recursive Filtering. *IEEE Trans. Geosci. Remote Sens.* **2014**, *52*, 3742–3752. [CrossRef]
59. Kang, X.; Xiang, X.; Li, S.; Benediktsson, J.A. PCA-Based Edge-Preserving Features for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 7140–7151. [CrossRef]
60. Duan, P.; Kang, X.; Li, S.; Ghamisi, P.; Benediktsson, J.A. Fusion of Multiple Edge-Preserving Operations for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 10336–10349. [CrossRef]
61. Ben Hamida, A.; Benoit, A.; Lambert, P.; Ben Amar, C. 3-D Deep Learning Approach for Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 4420–4434. [CrossRef]
62. Zhao, Z.; Xu, X.; Li, S.; Plaza, A. Hyperspectral Image Classification Using Groupwise Separable Convolutional Vision Transformer Network. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5511817. [CrossRef]
63. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5518615. [CrossRef]
64. Sun, L.; Zhao, G.; Zheng, Y.; Wu, Z. Spectral–Spatial Feature Tokenization Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5522214. [CrossRef]
65. Sun, L.; Zhang, H.; Zheng, Y.; Wu, Z.; Ye, Z.; Zhao, H. MASSFormer: Memory-Augmented Spectral-Spatial Transformer for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5516415. [CrossRef]
66. Mei, S.; Song, C.; Ma, M.; Xu, F. Hyperspectral Image Classification Using Group-Aware Hierarchical Transformer. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5539014. [CrossRef]
67. Wang, Z.; Li, C.; Xu, H.; Zhu, X. Mamba YOLO: SSMs-Based YOLO For Object Detection. *arXiv* **2024**, arXiv:2406.05835.
68. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: A Simple and Strong Anchor-Free Object Detector. 2020. Available online: https://arxiv.org/abs/2006.09214 (accessed on 12 October 2020).
69. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
70. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *arXiv* **2017**, arXiv:1708.02002.
71. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
72. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.