*remote sensing*

MDPI

*Article*

# AgeDETR: Attention-Guided Efficient DETR for Space Target Detection

**Xiaojuan Wang [1,2], Bobo Xi [1,3], Haitao Xu [1], Tie Zheng [1] and Changbin Xue [1,*]**

[1] National Space Science Center, Chinese Academy of Sciences, Beijing 100190, China; wangxiaojuan21@mails.ucas.ac.cn (X.W.); xibobo@xidian.edu.cn (B.X.); xuhaitao@nssc.ac.cn (H.X.); zhengtie@nssc.ac.cn (T.Z.)
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Telecommunications Engineering, Xidian University, Xi'an 710071, China
* Correspondence: xuechangbin@nssc.ac.cn

**Abstract:** Recent advancements in space exploration technology have significantly increased the number of diverse satellites in orbit. This surge in space-related information has posed considerable challenges in developing space target surveillance and situational awareness systems. However, existing detection algorithms face obstacles such as complex space backgrounds, varying illumination conditions, and diverse target sizes. To address these challenges, we propose an innovative end-to-end Attention-Guided Encoder DETR (AgeDETR) model, since artificial intelligence technology has progressed swiftly in recent years. Specifically, AgeDETR integrates Efficient Multi-Scale Attention (EMA) Enhanced FasterNet block (EF-Block) within a ResNet18 (EF-ResNet18) backbone. This integration enhances feature extraction and computational efficiency, providing a robust foundation for accurately identifying space targets. Additionally, we introduce the Attention-Guided Feature Enhancement (AGFE) module, which leverages self-attention and channel attention mechanisms to effectively extract and reinforce salient target features. Furthermore, the Attention-Guided Feature Fusion (AGFF) module optimizes multi-scale feature integration and produces highly expressive feature representations, which significantly improves recognition accuracy. The proposed AgeDETR framework achieves outstanding performance metrics, i.e., 97.9% in $mAP_{0.5}$ and 85.2% in $mAP_{0.5:0.95}$, on the SPARK2022 dataset, outperforming existing detectors and demonstrating superior performance in space target detection.

**Keywords:** space target detection; attention-guided feature enhancement; attention-guided feature fusion



**Citation:** Wang, X.; Xi, B.; Xu, H.; Zheng, T.; Xue, C. AgeDETR: Attention-Guided Efficient DETR for Space Target Detection. *Remote Sens.* **2024**, *16*, 3452. https://doi.org/10.3390/rs16183452

Academic Editor: Farid Melgani

Received: 3 August 2024
Revised: 4 September 2024
Accepted: 13 September 2024
Published: 18 September 2024



## 1. Introduction

With the rapid advancement of technology, the number of active spacecraft in orbit continues to increase [1]. Satellite infrastructure is crucial in various fields [2–6], such as communications, transportation, and weather forecasting, and has become indispensable in our daily lives. Ensuring the safety of space assets is of paramount importance [7–9], as space collisions pose a significant threat. This highlights the necessity for spacecraft that can autonomously detect surrounding objects, a capability crucial to reducing collision risks and enhancing Space Situational Awareness (SSA), which is key to maintaining the safety and sustainability of space operations. Since the 1960s, both the United States and the former Soviet Union have developed space surveillance systems [10]. These systems include ground-based and space-based detection methods and are now mainstream in monitoring space targets. The accurate identification of these targets from images captured by such systems is fundamental for effective space surveillance missions. As space technology evolves, detection techniques must also innovate to meet increasing monitoring demands and navigate more complex space environments. This evolution is vital to managing space traffic, safeguarding space assets, and promoting peaceful activities in space [11].

However, the challenges of complex space backgrounds, varying illumination conditions, and diverse target sizes significantly complicate the detection of space targets. Traditional methods often rely on filtering techniques, including spatial-domain [12,13], time-domain [14,15], and combined spatial–time-domain approaches [16,17]. Meng et al. [12] developed an adaptive technique for detecting dim and small objects against complex backgrounds. Their approach leverages spatial-domain filtering to enhance the contrast between target signals and background noise. Despite its effectiveness, the method exhibits sensitivity to parameter adjustments and shows limited robustness when dealing with strong noise or significant background variations. Smith et al. [15] introduced a temporal filtering method to improve space target detection by enhancing the signal-to-noise ratio. While this technique offers improved accuracy, it faces challenges in adapting to varying or unpredictable signal conditions due to its reliance on specific temporal patterns. Liu et al. [16] developed a spatio-temporal filtering strategy for detecting dim and small targets within aerospace systems. By integrating spatial and temporal filtering, their approach achieves better noise isolation and detection precision. However, this approach introduces significant computational complexity and sensitivity to parameter selection, which poses challenges for practical implementation. Overall, while these image processing methods are effective to an extent, they can be complex and typically produce only limited, low-level visual features which do not meet the rigorous requirements for space target detection. In particular, against the intricate backdrop of a starry sky, dim and faint targets can easily be obscured by background noise, presenting a significant challenge for conventional filtering methods.

With the rise of deep learning, coupled with rapid advancements in hardware computing power, significant progress has been made in computer vision, particularly in image classification [18,19], segmentation [20], and object detection [21]. Recent research has increasingly focused on optimizing and enhancing space target detection algorithms through deep learning. With its rapid development, target detection technology in computer vision has gradually outperformed traditional image processing methods [22,23]. A notable advantage of deep learning-based target detection is its ability to identify space targets effectively under challenging conditions. For instance, Xue et al. [24] developed a Convolutional Neural Network (CNN) model specifically designed for detecting weak and small targets against starry sky backgrounds and achieved pixel-level segmentation. Similarly, Xiang et al. [25] employed a Fast Grid-Based Neural Network architecture, which divides images into $14 \times 14$ segments to locate space debris accurately. Additionally, Xi et al. [26] proposed a comprehensive detection framework that combines pre-processing techniques to generate candidate regions for classification using CNNs, thus producing robust detection outcomes. These CNN-based methods illustrate the dominance of such systems in the target detection field. Among these, You Only Look Once (YOLO) detectors [27–34] are notable for their superior performance in detection accuracy and speed. However, these detectors typically need Non-Maximum Suppression (NMS) for post-processing, which increases the computational load and may hinder real-time performance. Therefore, it is crucial to develop algorithms that function independently of NMS to streamline detection processes and improve operational efficiency in target detection.

Subsequently, Carion et al. [35] introduced Detection Transformer (DETR), an end-to-end target detection framework that quickly gained significant attention in the academic community. DETR stands out by providing an end-to-end training method that eliminates the need for traditional NMS post-processing and significantly enhances the understanding of the image context through a self-attention mechanism inspired by Transformers. However, despite its considerable promise in target detection tasks, DETR encounters challenges such as high sensitivity to parameters, computational intensity, and limitations in handling multi-scale targets. These issues may constrain its real-time performance and broad applicability. To address these issues, researchers have suggested various enhancements, including optimizing query initialization [36–40], adjusting the attention architecture [41–44], and effectively utilizing multi-scale features [45–47].

For instance, Zhang et al. introduced the DINO [40] model, which features a significant advancement by employing a novel mixed query selection strategy. The strategy is utilized by the encoder output to initialize the decoder position queries while preserving the learnability of content queries. This method enables the effective use of enhanced positional information and significantly improves the ability of the model to aggregate and utilize comprehensive content features from the encoder. Another key innovation was the dual-query mechanism in DQ-DETR [37], which typically relies on a fixed set of queries for predicting target bounding boxes and categories. Thus, this approach notably improves the target detection capabilities of the model compared with traditional models like DETR. Deformable DETR [41] presents a deformable attention mechanism that improves computational efficiency and facilitates the use of multi-scale feature maps for predictions. While Deformable DETR is more efficient, its deformable attention mechanism is more complex to implement than standard Transformer attention. Dynamic DETR [44] further enhances adaptability and performance in target detection tasks through dynamic adjustment mechanisms. These dynamic features allow the model to respond more flexibly to variations in input conditions, thus enabling efficient and accurate target recognition and localization. Cao et al. also contributed to the field, with CF-DETR [45], a model that effectively harnesses multi-scale features to enhance spatial perception within its detection framework. This approach significantly improves precision and efficiency in detecting small and less distinguishable targets.

In this study, we refine the DETR model specifically for space target detection due to its substantial advantages over YOLO detectors. The end-to-end detection framework of DETR enhances precision and robustness in complex spatial environments through improved context understanding. Unlike YOLO detectors, DETR models eliminate traditional post-processing steps such as NMS, which simplifies the detection process and potentially increases detection speed. However, despite their strong performance in natural image detection, DETR models have limitations in optimization for the specific characteristics of space targets, which affects their practical application performance in space target detection. To tackle this issue, we conducted comprehensive analysis and evaluation of space target characteristics, leading to the development of AgeDETR. This novel model builds on the strengths of DETR while introducing significant improvements by simplifying the network structure and incorporating advanced attention mechanisms. These enhancements delve into the advantages of advanced attention mechanisms, providing a deeper understanding and addressing the specific needs of space target detection. Experimental evaluations on the SPARK2022 [48] public dataset show that AgeDETR achieves exceptional target detection performance, with 97.9% in $mAP_{0.5}$ and 85.2% in $mAP_{0.5:0.95}$.

AgeDETR introduces several significant advancements to enhance the detection of space targets by incorporating sophisticated attention mechanisms and advanced network architectures. This section outlines the key contributions of AgeDETR and details its novel approach and technological improvements:

1. We introduce AgeDETR for space target detection, which significantly improves detection performance by incorporating advanced attention mechanisms into the backbone network and encoder. The model comprises four principal components: the EF-ResNet18 backbone network, the AGFE module, the AGFF module, and a sophisticated decoder that ensures precise target localization and classification. Together, these components effectively tackle the common challenges in space target detection.
2. To tackle illumination variability encountered in space target detection, we design the EF-ResNet18 architecture as the backbone network. This architecture creatively combines the FasterNet block [49] and EMA [50] technologies to establish the EF-Block module, which is seamlessly integrated into ResNet18 [51]. This design significantly boosts the feature extraction capabilities of the backbone network and optimizes computational efficiency. With these enhancements, EF-ResNet18 provides stable target detection under varying lighting conditions and improves the precision and robustness of the detection results.

3. To overcome the challenges posed by complex space backgrounds, we propose the AGFE module. This module meticulously integrates two complementary attention mechanisms, specifically designed to enhance target feature recognition and optimize the extraction of critical information. The AGFE module also employs a single-layer Transformer encoder to efficiently process high-level features, simplifying computational steps. This strategy significantly improves the accuracy of the model in recognizing and locating targets against complex backgrounds and optimizes computational efficiency.
4. To address the issues associated with diverse target sizes, we introduce the AGFF module. Unlike traditional multi-scale fusion methods for natural images, AGFF employs an attention-guided strategy. This strategy facilitates the fusion of features across adjacent layers through high-level attention mechanisms, which effectively enhance feature fusion performance. Consequently, this module significantly enhances the capability of the model to detect and classify targets of varying scales accurately, thereby boosting overall detection performance.

The following sections of this paper are organized as follows: Section 2 provides a brief overview of related work in the field. Section 3 details the methodology of AgeDETR. Section 4 presents experimental results and analyses. Section 5 concludes the paper by summarizing the key findings.

## 2. Related Work

In space target detection, the evolution of object detection algorithms and the integration of attention mechanisms are paramount. This section explores the current state of object detection algorithms, with a focus on the advancements and limitations of prominent methods such as YOLO and DETR. Additionally, we examine the critical role of attention mechanisms in enhancing detection accuracy and efficiency. By analyzing these technologies, we aim to provide effective solutions for space target detection.

### *2.1. Current State of Object Detection Algorithms*

Space target detection requires high accuracy and real-time performance, making it essential to select an optimal detection method that effectively balances these requirements. YOLO detectors have garnered widespread acceptance in this area due to their efficient and precise single-stage object detection capabilities. Since Joseph Redmon first introduced YOLOv1 [27] in 2015, the algorithm has evolved through multiple iterations, with each one breaking through previous limitations and enhancing detection performance. YOLOv1 faced limitations in handling objects of varying sizes due to its reliance on fully connected layers, which restricted its ability to generalize across different scales. YOLOv2 [28] addressed this by incorporating anchor boxes, which facilitated the automatic determination of optimal anchor sizes, but still struggled with complex backgrounds. YOLOv3 [29], with its deeper Darknet-53 network, improved feature extraction but encountered challenges in balancing speed and accuracy. YOLOv4 [52] introduced the CSPDarknet53 backbone network, refining feature fusion strategies to improve the detection of smaller objects, yet still required further enhancement in handling low-contrast scenarios. YOLOv7 [32] introduced the Convolutional Block Attention Module (CBAM) [53], which significantly enhanced the feature extraction capabilities by focusing on relevant parts of the image, though it added some computational overhead. YOLOv8 [33], with its Cross Stage Partial Network with Two Filters construction blocks, continued to optimize the network architecture but faced difficulties in maintaining real-time performance for very-high-resolution images. YOLOv9 [34] addressed limitations in detection accuracy through advancements in the loss function, which improved the precision in detecting objects under challenging conditions, though the model complexity slightly increased.

Despite the successful application of YOLO detectors in general object detection tasks, their reliance on CNN architectures may restrict their ability to capture the comprehensive global context. Additionally, these detectors typically depend on NMS for post-processing

to eliminate overlapping predictions and improve accuracy. However, implementing NMS is intricate, and its hyperparameters significantly affect performance, which can potentially compromise detection speed and robustness. Consequently, developing algorithms that operate independently of NMS is essential to streamlining the detection process and enhancing operational efficiency. In contrast, DETR, proposed by Facebook AI in 2020, has made significant advancements in object detection tasks. DETR simplified the detection process by predicting sets of objects directly without the need for candidate region extraction or NMS post-processing. By leveraging the self-attention mechanism of the Transformer architecture, DETR has significantly enhanced the understanding of image context and achieved breakthrough performance in object detection tasks.

The decision to investigate DETR for space target detection in this study stems from its inherent advantages over YOLO detectors. While YOLO excels in real-time performance, DETR provides superior context understanding and eliminates the complexities associated with NMS. However, DETR faces challenges such as parameter sensitivity, high computational requirements, and limitations in handling multi-scale objects. To address these challenges, a series of improvements have been proposed. For instance, while Deformable DETR introduced a deformable attention module that enhanced image feature processing and accelerated model training, it also had to manage the increased complexity introduced by deformable attention. Efficient DETR [54] improved the initialization process of object queries and reference points to accelerate model convergence but required the careful tuning of these parameters to maintain performance. Dynamic DETR, despite enhancing model performance and training efficiency with a dynamic attention mechanism, had to overcome the challenge of balancing model complexity with efficiency gains. Lite DETR [55] optimized the efficiency of the encoder by reducing the update of low-level features, though this reduction necessitated additional strategies to preserve detection accuracy. Anchor DETR [36] improved interpretability with an anchor-based query design and reduced memory consumption with the Row–Column Decoupled Attention variant; however, this approach required careful management of anchor sizes to avoid potential issues in detection precision. Conditional DETR [38] increased positioning accuracy and expedited the training process with conditional spatial queries, but at the cost of increased complexity in the query design. DAB-DETR [56] adopted four-dimensional dynamic anchor boxes, using prior location information to accelerate model convergence, though it introduced additional overhead in managing dynamic anchors. DN-DETR [39] introduced noisy object queries to reduce instability in the matching process, effectively solving the slow convergence issue but necessitating careful noise management to avoid degradation in performance. RT-DETR [47] integrated the advantages of Transformer and DETR to achieve accurate real-time object detection, though balancing real-time performance with detection accuracy remained a challenge. In this paper, we build upon the strengths of DETR to develop a robust and efficient model specifically designed for space target detection. By focusing on reducing computational complexity and optimizing the architecture, we propose a tailored solution that effectively addresses the unique challenges of this task.

### *2.2. Attention Mechanism*

Attention mechanisms are pivotal in computer vision and greatly enhance the efficiency of visual information processing in models. They improve the interpretation of complex scenes and increase the adaptability of systems across diverse visual tasks. The diversity and flexibility of attention mechanisms allow for adaptation to various visual tasks and data types. The main types of attention mechanisms include the following:

- Channel attention. Channel attention boosts performance by dynamically adjusting the importance of each channel and selectively focusing on relevant features. Hu et al. [57] introduced the concept of channel attention with the development of SENet, centered around the Squeeze-and-Excitation (SE) block. This block collects global information, captures channel-wise relationships, and enhances representational capabilities. It recalibrates channel-wise feature responses to improve feature

discriminability. Nevertheless, an SE block captures global information solely through global average pooling, which restricts its modeling capability, limiting its ability to capture complex interactions between channels. To address this issue, Gao et al. [58] introduced a Global Second-order Pooling block into the squeeze module. This module enables the modeling of high-order statistics and the synthesis of global information, thus enhancing the expressive power of the network. However, this enhancement increases computational demands. Lee et al. [59] developed the lightweight Style-based Recalibration Module (SRM) to overcome the limitations of existing channel attention methods. This module effectively recalibrates CNN feature maps by extracting and integrating style information from each channel. SRM utilizes style pooling to derive style details from the channels and assigns recalibration weights through channel-independent integration. Despite its effectiveness, the focus on style information might not generalize well to tasks requiring a broader context. Wang et al. [60] introduced the Efficient Channel Attention (ECA) block, which uses a 1D convolution to determine the interaction between channels instead of relying on dimensionality reduction. This approach significantly reduces the computational complexity associated with dimensionality reduction techniques, making the ECA block more efficient. Still, while this method improves computational efficiency, it also introduces challenges in capturing more complex channel interactions that might be better addressed with higher-dimensional convolutions or more sophisticated attention mechanisms.

- Spatial attention. Spatial attention is a mechanism that adaptively selects and emphasizes specific spatial regions. By focusing on areas of interest, this approach optimizes the extraction of relevant features and improves overall performance. For instance, Mnih et al. [61] proposed the Recurrent Attention Model (RAM), which sequentially focuses on different regions. This approach enables the model to process one part of the input at a time and decide on the subsequent focus point, mirroring the human method of scanning visual scenes. Although the RAM demonstrates effectiveness in tasks requiring sequential attention and context accumulation, it relies heavily on sequential processing, which might limit its application in real-time tasks. From another perspective, CNNs excel at processing image data due to their translation equivariance. Nonetheless, they lack rotation, scaling, and warping invariance, limiting their robustness in certain scenarios. To address these limitations, Jaderberg et al. [62] proposed Spatial Transformer Networks (STNs), the first attention mechanism explicitly designed to predict relevant regions and provide transformation invariance to deep neural networks. While STNs enhance the ability to focus on important regions and learn these invariances, they also introduce increased model complexity and potential computational overhead. The limitations of both the RAM and STNs suggest that while each addresses specific challenges, neither fully overcomes the trade-offs among efficiency, accuracy, and complexity.
- Hybrid attention. Hybrid attention integrates channel and spatial attention mechanisms for a holistic understanding of image features. Woo et al. [53] proposed a hybrid attention mechanism known as CBAM. By sequentially combining channel and spatial attention, CBAM leverages the spatial and cross-channel relationships of features to guide the network on what and where to focus. Despite its effectiveness in enhancing feature selection, CBAM faces the challenge of increased model complexity. The Residual Attention Network [63] highlights the importance of informative features across spatial and channel dimensions. This network uses a bottom-up structure with multi-level convolutional layers to generate a three-dimensional attention map encompassing height, width, and channel. However, it faces challenges such as high computational expenses and limited receptive field expansion. To address these challenges, Park et al. [64] proposed the Bottleneck Attention Module (BAM) to enhance network representational capability. The BAM uses dilated convolutions to broaden the receptive field and implements a bottleneck structure, which helps to minimize computational expenses. Additionally, it adjusts features in both the

channel and spatial dimensions, thus enhancing feature representation. Despite the effectiveness of dilated convolutions in expanding the receptive field, they struggle to capture long-range contextual information and enhance cross-channel relationships. Liu et al. [65] proposed Cross-scale Attention, a mechanism that enables dynamic feature interaction across different scales. By integrating information from multiple scales, this approach enhances the robustness of feature representation, allowing the model to more effectively address challenges arising from scale variations. However, the process of fusing features across scales could introduce additional computational complexity, potentially increasing inference time. Ouyang et al. proposed EMA, a multi-scale attention mechanism designed to effectively focus on relevant features across different scales. This mechanism aims to retain information within each channel while reducing computational overhead. By dynamically allocating attention to various scales, EMA enhances model performance. It integrates feature information from different levels, enabling the model to detect local details while perceiving global context. This fusion of multi-scale features has proven crucial to advancing computer vision research, particularly in handling complex visual scenes.

- Self-attention mechanism. The self-attention mechanism, initially introduced by Vaswani et al. [66] in natural language processing, has significantly impacted computer vision, especially in object detection. This mechanism is highly effective in capturing long-range dependencies and the global context, which is essential to comprehending complex visual scenes. Specifically, Vision Transformer (ViT) presented by Dosovitskiy et al. [67] demonstrates the power of self-attention in computer vision. By treating image patches as tokens and applying Transformer to these sequences, ViT achieves competitive performance compared with traditional CNNs on large-scale image classification tasks. Even so, ViT encounters challenges in processing high-resolution images due to its reliance on a fixed number of tokens, which limits its ability to efficiently handle finer details. Additionally, Wang et al. [68] developed Non-Local Neural Networks, which apply self-attention in video and image tasks to capture long-range dependencies more effectively than convolutional layers. This method has been shown to have greater performance on various tasks, including video classification and object detection, by incorporating global information into the feature representation. Despite these improvements, Non-Local Neural Networks still face challenges in computational efficiency due to the quadratic complexity of the self-attention mechanism. Particularly, Carion et al. pioneered using self-attention in object detection with the introduction of DETR. This model uses the Transformer architecture to process entire images as sequences, directly modeling relationships between distant regions. Although DETR achieves state-of-the-art performance, it struggles with slow convergence during training and requires large datasets to generalize effectively. Building on the success of DETR, Zhu et al. introduced Deformable DETR, which improves the original with deformable attention modules. These modules dynamically adjust the receptive fields based on the input features, making the model more efficient with high-resolution images and improving detection accuracy. Nevertheless, while Deformable DETR improves efficiency, it still faces challenges in balancing complexity with accuracy, especially in scenarios requiring real-time processing. Over the years, significant improvements have been made to enhance the efficiency and effectiveness of self-attention mechanisms. For instance, Yin et al. [69] introduced Disentangled Non-Local Neural Networks, which disentangle non-local operations to boost the representational power of self-attention. This enhancement enables more accurate and effective capture of long-range dependencies. Even with these advancements, challenges remain in optimizing the trade-off between computational cost and the accuracy of long-range dependency modeling.

In this study, we adopt attention mechanisms to enhance the performance of models in space target detection. These mechanisms allow a model to selectively highlight key features when processing space images, thereby improving detection accuracy and stability

in dynamic and complex environments. We chose to apply attention-guided strategies because they efficiently capture local and global feature dependencies, which are essential to accurately identifying and classifying space targets under varying conditions.

## 3. Method

This section provides a comprehensive overview of our approach. Section 3.1 introduces the enhanced backbone network, EF-ResNet18, Sections 3.2 and 3.3 provide detailed explanations of the AGFE and AGFF modules, respectively.

AgeDETR, depicted in Figure 1, consists of four fundamental components: EF-ResNet18, AGFE, AGFF, and a decoder. The EF-ResNet18 backbone integrates the EF-Block module, which combines the advanced FasterNet block architecture with the EMA mechanism. This integration significantly enhances the performance of the traditional ResNet18 by optimizing computational workflows and boosting feature extraction efficiency for space targets. The EF-Block module allows the backbone to manage varying illumination conditions inherent in space target detection by dynamically adjusting feature extraction processes, which ensures more accurate and reliable detection performance. The AGFE module incorporates self-attention and channel attention mechanisms across multiple layers of the backbone network. This dual attention approach allows the model to discern and prioritize essential features, which ensures that the critical aspects of space targets are highlighted. By focusing on these key features, the AGFE module optimizes the extraction of space target characteristics, effectively mitigating the challenges posed by complex space backgrounds. This refined feature extraction enhances the model in distinguishing space targets from noise and significantly boosts overall detection accuracy and robustness. The AGFF module facilitates the transfer of high-level features from the Transformer encoder to lower levels. By utilizing attention-weighted feature fusion, AGFF implements a top-down strategy to create a new, more expressive feature layer. This method ensures efficient multi-scale feature fusion, which is crucial to accurately identifying and localizing space targets of varying sizes. By integrating features across different scales, the AGFF module detects small and large targets more effectively, which improves detection precision and robustness. The decoder, equipped with auxiliary prediction heads, iteratively refines object queries and generates precise predictions for object categories and their bounding boxes.

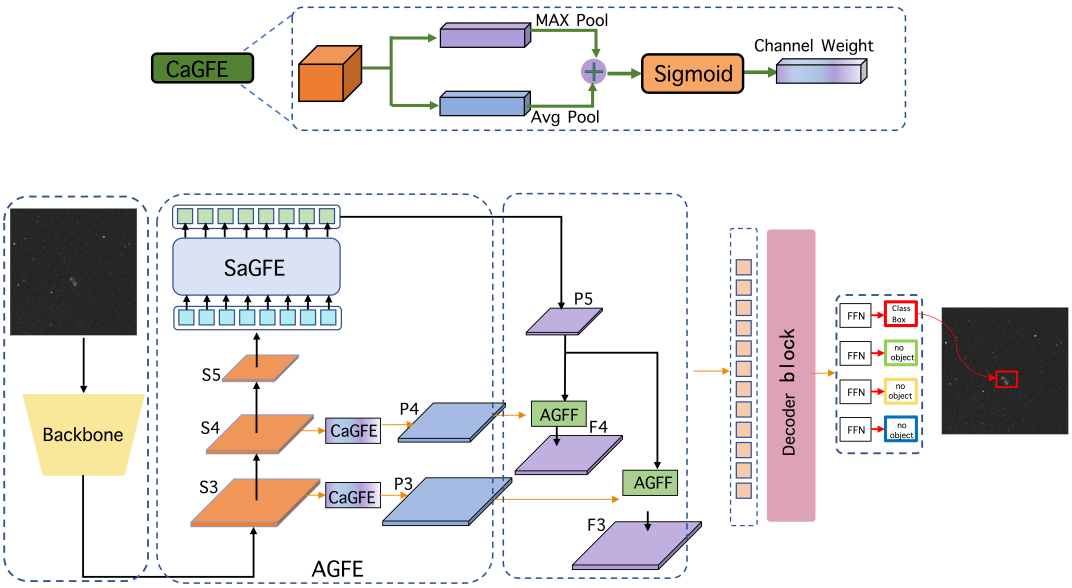



**Figure 1.** Overview of the proposed AgeDETR.

### *3.1. EF-ResNet18*

In the EF-ResNet18 architecture, EF-Block (Figure 2c) is crucial, as it integrates innovations from the FasterNet block and incorporates EMA. The FasterNet block introduces efficient convolutional operations that accelerate processing speed while maintaining accuracy. The EMA mechanism allows the model to focus on relevant features across different scales, improving the robustness and precision of feature extraction. This integration reduces model parameters, lowers computational demand, and improves the capability to process multi-channel information. These advancements are vital for the robust recognition of space target characteristics under varying illumination conditions, ensuring improved performance and technical reliability.

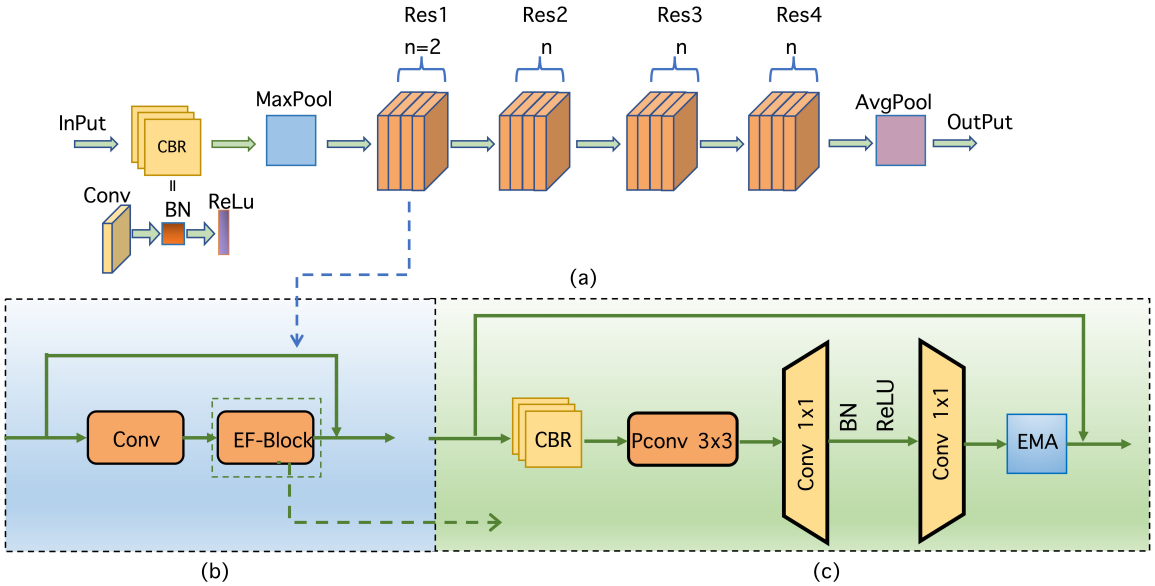



**Figure 2.** (**a**) Overview of the overall framework of the proposed EF-ResNet18 architecture. (**b**) The improved residual connection structure incorporating the EF-Block module. (**c**) A detailed structural diagram of the EF-Block module.

EF-Block employs the FasterNet block, which begins with a Partial Convolution (PConv) layer [49], followed by two Pointwise Convolution (PWConv) layers. The PConv layer serves as the core of the FasterNet block and introduces a novel convolution approach that promotes feature extraction efficiency while reducing computational redundancy. As shown in Figure 3, PConv selectively applies filters to specific input channels, leaving others unmodified. Assuming that the input and output feature maps both have $C$ channels, the floating-point operations (FLOPs) for PConv can be expressed as

$$h \times w \times k^2 \times C_p^2 \tag{1}$$

where $h$ and $w$ are the dimensions of the output feature map, $k$ is the kernel size, and $C_p$ denotes the number of channels involved in the convolution. In contrast, the FLOPs for regular convolution are

$$h \times w \times k^2 \times C^2 \tag{2}$$

by defining the ratio $r = \frac{C_p}{C}$, the FLOPs of PConv amount to only $r^2$ of those for a regular convolution. This selective filtering significantly enhances performance by avoiding uniform processing across all channels. The subsequent PWConv layers further refine and consolidate the extracted features. EF-Block fuses the FasterNet block with the EMA module. The integration of the FasterNet block optimizes convolution operations by directing computational resources to relevant channels, which improves both speed and efficiency. Additionally, the EMA module encodes global information to recalibrate channel weights

across parallel branches, aggregates output features through cross-dimensional interaction, and captures crucial pixel-level relationships. By prioritizing the most relevant features and enabling comprehensive feature aggregation from multiple dimensions, the module captures essential spatial details and relationships. This is vital to accurately detecting and analyzing complex features, particularly under varying illumination conditions. The integration boosts the capability to represent features more effectively, enhancing both the accuracy and efficiency of feature extraction. In the ResNet18 architecture, each residual block consists of two convolutional layers: the first layer is responsible for spatial feature extraction, while the second layer typically refines the features and reduces dimensionality. In this study, we seamlessly integrate EF-Block into the residual network framework of ResNet18 by replacing the second standard convolutional layer in each residual block with our module (as shown in Figure 2a). This approach enhances overall performance and efficiency.

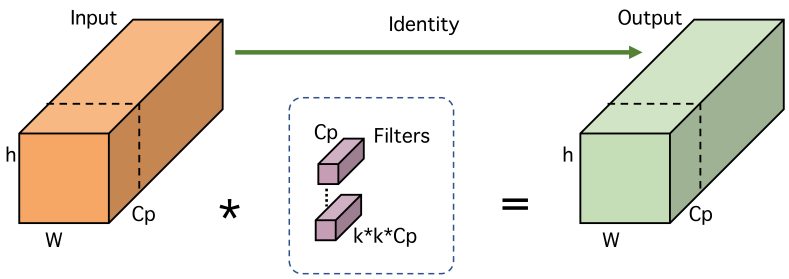



**Figure 3.** The schematic of the principle of Partial Convolution.

In summary, the EF-ResNet18 architecture represents an advanced approach to feature extraction and representation, leveraging the advantages of the EF-Block module within the ResNet18 framework. This architecture enhances convolution operations through selective filtering and global information encoding, significantly improving the speed and precision of feature handling. These methodological improvements are critical for robust performance in environments with complex spatial dynamics.

### *3.2. AGFE Module*

We incorporate EF-ResNet18 as the backbone of AgeDETR to leverage its outstanding feature extraction capabilities. However, despite its effectiveness, the CNN-based architecture of EF-ResNet18 faces challenges in capturing comprehensive global features of space targets, which can impact target localization precision and classification accuracy. To overcome this limitation, we integrate the feature sets from the final three stages of the backbone $\{S3, S4, S5\}$ into an advanced AGFE module, as depicted in Figure 1. This module strengthens feature integration and representation by enhancing spatial coherence and contextual understanding. The AGFE module comprises two complementary components that work in tandem to enhance the performance of the model. The following section delves into the specifics of these two attention mechanisms.

#### 3.2.1. Self-Attention-Guided Feature Enhancement (SaGFE) Module

We propose the SaGFE module, a component specifically designed to refine the feature extraction process, with a particular focus on enhancing the expressiveness and computational efficiency of high-level features denoted by $S5$. The SaGFE module is applied specifically to the $S5$ feature layer because incorporating self-attention mechanisms at this level enables the precise capture and association of rich semantic concepts, which is crucial for accurate object localization and recognition in subsequent modules. In contrast, applying the SaGFE module to low-level features would likely result in redundant processing and potential confusion due to their limited semantic information, thereby reducing overall model performance. The module integrates a single-layer Transformer encoder that applies self-attention directly to $S5$ features to dynamically focus on the most relevant aspects of

the input without traditional sequential constraints. The computational steps involved are detailed in the following equations:

$$Q = K = V = Flatten(S5) \tag{3}$$

$$P5 = Reshape(SaGFE(Q, K, V)) \tag{4}$$

where $S5$ is flattened into a one-dimensional array, which then serves simultaneously as the $Query(Q)$, $Key(K)$, and $Value(V)$ inputs, essential for the self-attention mechanism. The SaGFE module incorporates a self-attention function specifically designed to capture global information from these inputs efficiently. The resulting data are subsequently reshaped into $P5$, ensuring compatibility with subsequent layers and facilitating integration into the overall network architecture.

The SaGFE module stands out due to its innovative use of a single-layer Transformer encoder, which fundamentally changes how high-level features are processed in space target identification and localization. Unlike traditional CNNs, which focus on local features and often struggle to capture long-range dependencies, the module dynamically evaluates the entire feature set, uncovering subtle and distant relationships crucial for a comprehensive understanding of space targets. This capability enables the module to prioritize the most relevant aspects of the input data, free from the constraints of traditional sequential processing. As a result, it adapts more effectively to diverse and unpredictable environments. Additionally, this approach reduces computational demands by utilizing a single-layer encoder. By deeply integrating the advantages of Transformer technology into feature extraction, the SaGFE module provides a more nuanced, efficient, and precise analysis of space targets.

#### 3.2.2. Channel Attention-Guided Feature Enhancement (CaGFE) Module

We propose the CaGFE module to process low-level features $S3$ and $S4$. The framework of the CaGFE module is illustrated in Figure 2. Initially, the CaGFE module processes the input feature map $f_{\text{in}} \in \mathbb{R}^{C \times H \times W}$, where $C$ represents the number of channels, and $H$ and $W$ denote the height and the width of the feature map, respectively. The CaGFE module harnesses the strengths of global average and global maximum pooling by applying these operations sequentially to $f_{\text{in}}$. This synergistic combination incorporates global context and local details, which enriches the overall feature expression. Mathematically, this process can be represented as

$$CA = \delta(f_{c1}(MaxPool(x)) + f_{c2}(AvgPool(x))) \tag{5}$$

$$P_i = CA(S_i) * S_i \tag{6}$$

where $CA$ represents the channel attention function, $\delta$ denotes the sigmoid function, and $i$ indicates the index of pyramid levels. The $CA$ mechanism is vital to boosting performance by learning channel interdependencies and dynamically adjusting their weights. This process emphasizes crucial features while suppressing irrelevant ones. By focusing on lower-level features like $S3$ and $S4$, rich in spatial details, the $CA$ mechanism improves recognition accuracy. It strengthens important intra-channel features, enhancing overall discriminative power. Additionally, the $CA$ module is lightweight and efficient, enhancing low-level features with minimal computational overhead, thereby preserving efficiency.

The AGFE module synergistically enhances feature integration and representation by combining the SaGFE and CaGFE components. The SaGFE module, with its innovative use of a single-layer Transformer encoder, excels at capturing long-range dependencies and the global context, while the CaGFE module focuses on enhancing low-level feature details. This integration significantly improves efficiency and the capability to detect and localize space targets, ensuring exceptional performance even in challenging environments with complex backgrounds.

### *3.3. AGFF Module*

In deep learning networks, hierarchical feature extraction reveals a wealth of information. Low-level features capture fundamental visual elements such as edges and textures, which are crucial for spatial resolution and local detail expression. In contrast, high-level features encapsulate sophisticated semantic attributes and offer critical insights for a comprehensive understanding of image content. Leveraging this principle, fusing features across different levels of abstraction significantly enhances the comprehensiveness of feature representation and achieves a synergistic effect that combines the strengths of each level. The multi-level feature integration strategy is pivotal to improving the accuracy and robustness of space target detection. A common approach involves directly adding upsampled high-level and low-level features pixel-wise. However, this method lacks a strategic feature selection process and merely sums pixel values across layers without discrimination. To address this limitation and fully utilize the potential of multi-scale feature fusion, we introduce the AGFF module. This module optimizes the feature fusion process within the network by intelligently selecting and integrating the most informative features from adjacent layers, which enhances overall model performance. This improvement enables the model to detect space targets with greater precision and robustness across multiple scales.

In specific, AGFF integrates attention mechanisms with Feature Pyramid Networks [70] to enhance the precision of multi-scale target detection. By utilizing the attention weights generated by these high-level features as a key fusion factor for integrating low-level features, the AGFF module ensures that the fusion process is both strategic and effective. This approach allows for more precise selection and weighting of features, which significantly improves the ability to capture and synthesize critical information from diverse feature representations. The primary advantage of using attention weights is that they provide a dynamic and context-aware method for emphasizing the most relevant features, leading to richer and more accurate feature integration. This creative fusion approach ensures sophisticated weighting and consolidation of features across different levels, which improves the ability to capture and synthesize critical information from diverse feature representations. To maintain the spatial dimensional consistency of feature maps, we combine transposed convolution and bilinear interpolation techniques. The transposed convolution expands the feature map and reconstructs input through learned parameters, while bilinear interpolation offers a rapid and direct method for scale adjustment, effectively addressing non-uniform sampling issues. The integration of these technologies enhances performance and streamlines processing efficiency, which leads to an effective solution for multi-scale target detection. The formulations for this process are detailed in (7) and (8):

$$\alpha = CaGFF(BL(T - Conv(P_5))) \tag{7}$$

$$N_i = P_i * \alpha + P_{i+1} \tag{8}$$

where $T - Conv$ and $BL$ represent transposed convolution and bilinear interpolation, respectively. $\alpha$ serves as the fusion factor, and $i$ denotes the index of pyramid levels. Initially, we employ a combination of transposed convolution and bilinear interpolation techniques to adjust the scale of high-level features. After that, adjusted feature $P_5$, as shown in Figure 1, undergoes processing in the CaGFE module, which enhances these high-level features by using attention mechanisms and computes the fusion factor $\alpha$, a key parameter for feature fusion.

The AGFF module significantly enhances multi-scale target detection by optimizing feature fusion through attention mechanisms and integrating feature pyramids. This method improves precision and strengthens the capability of the model to capture and utilize critical information across different levels of feature abstraction. Consequently, the model can detect space targets with higher accuracy and robustness across various scales.

## 4. Experiments

### *4.1. Datasets and Evaluation Measures*

We conducted experiments on the SPARK2022 dataset [48], and some examples are shown in Figure 4. The SPARK2022 dataset is a unique space multi-modal annotated image dataset containing 110k RGB images of 11 object classes, including 10 spacecraft and 1 class of space debris. These images were generated in a realistic space simulation environment, encompassing diverse sensing conditions, such as extreme orbital scenarios, background noise, low signal-to-noise ratio, and high image contrast typical of space imagery. For our experiments, we randomly partitioned the SPARK2022 dataset into a training set of 8800 images and a validation set of 2200 images to evaluate our proposed algorithm. Figure 5 provides a detailed quantitative analysis of the dataset characteristics. Specifically, Figure 5a shows the exact instance counts for each object category, while Figure 5b depicts the distribution of bounding box sizes, illustrating the range of object dimensions within the dataset. Figure 5c illustrates the distribution of object–bounding box centers and highlights a concentration in the mid-areas of the images with darker tones. Finally, Figure 5d presents a scatter plot correlating their width and height, which underscores the dataset's significant scale diversity and the associated challenges in accurate space target detection. Through rigorous training and validation, we demonstrated the practicality and efficacy of our model in space target detection.

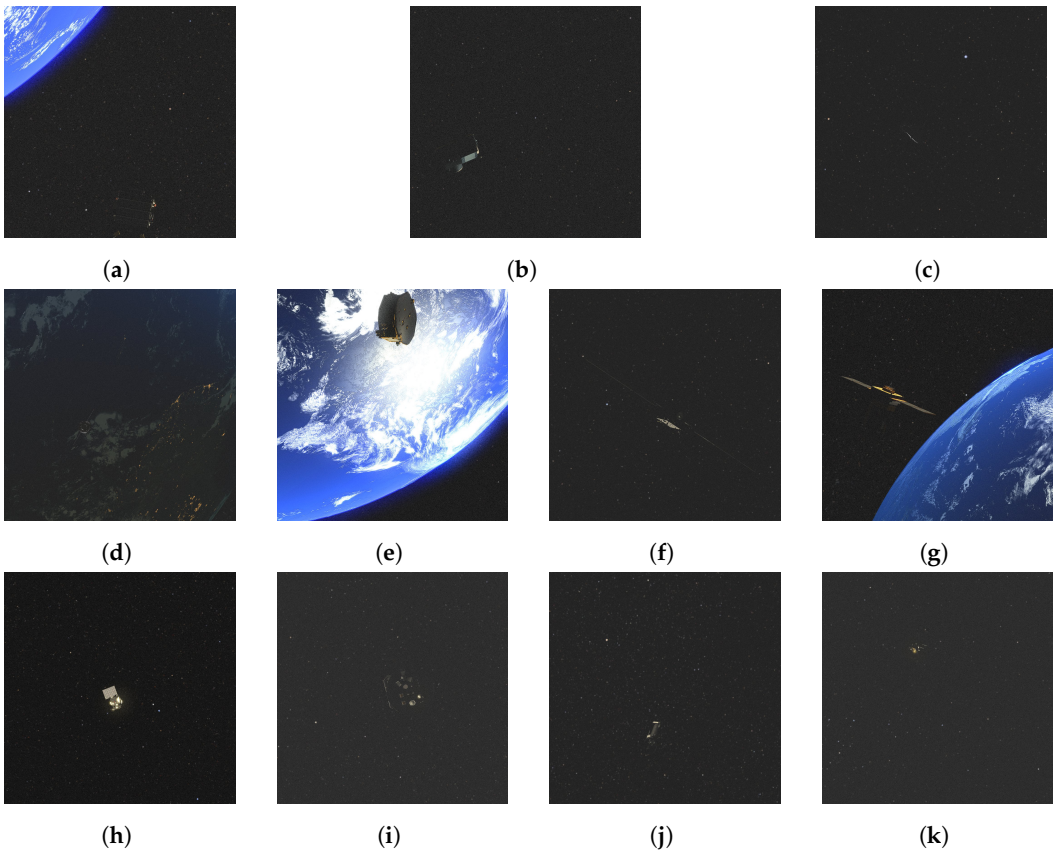



**Figure 4.** Examples of SPARK2022 dataset. (**a**) Proba 2, (**b**) Cheops, (**c**) Debris, (**d**) Double star, (**e**) Lisa Pathfinder, (**f**) Smart 1, (**g**) Soho, (**h**) Proba 3 CSC, (**i**) Proba 3 OCS, (**j**) Xmm newton, and (**k**) Earth Observation Sat 1.

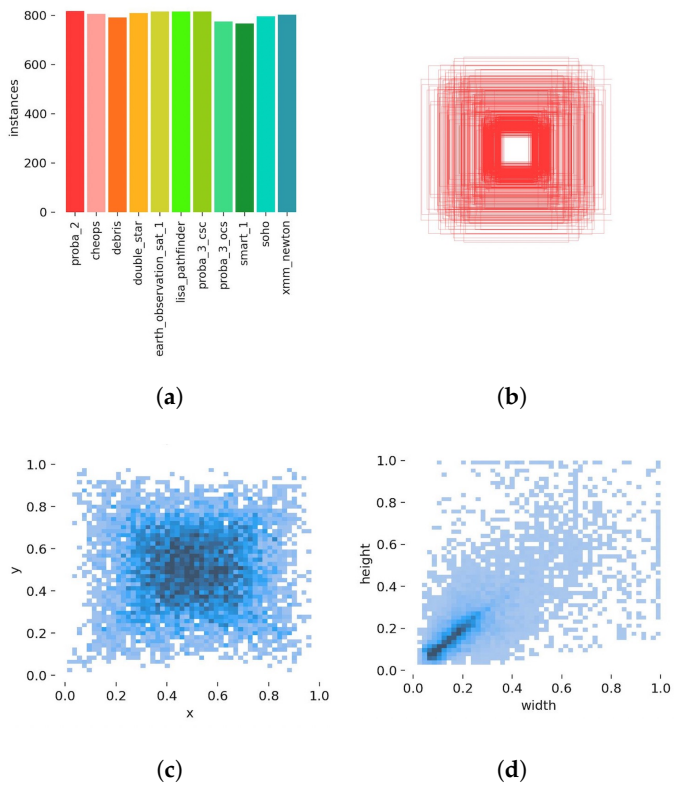



**Figure 5.** Information about the manual labeling of the objects in the SPARK2022 dataset.

To thoroughly assess the performance of our model, we utilized the standard evaluation metrics in object detection for a comprehensive performance analysis. These metrics include precision, recall, mean average precision $(mAP)$, the count of model parameters, and the computational complexity (GFLOPs). Precision measures the proportion of correctly identified targets among all detection results of the model, which is calculated as

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

where $TP$ denotes the true positives (the number of correctly identified targets) and $FP$ denotes the false positives (the number of targets incorrectly identified as positive). Recall reflects the ability of the model to detect all actual targets and is defined as

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

where $FN$ signifies false negatives (the number of targets that were not detected). Average precision $(AP)$ is a crucial metric that provides an aggregate measure of the performance of the model across different object categories. It evaluates the precision–recall curve and averages the precision values at different recall levels. Specifically, $mAP_{0.5}$ represents the mean average precision at an Intersection over Union $(IoU)$ threshold of 0.5, while $mAP_{0.5:0.95}$ indicates the mean $AP$ computed across an $IoU$ range from 0.5 to 0.95 with a step size of 0.05. These metrics offer a comprehensive evaluation of the ability of the model to detect objects accurately under various $IoU$ criteria, reflecting its robustness and reliability in diverse detection scenarios. Additionally, we considered the count of model parameters and computational complexity (measured in GFLOPs) to assess the efficiency and scalability of our model in practical deployment scenarios.

We used these metrics to perform a comprehensive analysis of the proposed model's performance in space target detection. This analysis demonstrated the effectiveness and suitability of the model for real-world applications.

### *4.2. Experimental Settings*

We developed the AgeDETR model by utilizing Python and the PyTorch deep learning framework. To initialize the object queries in the decoder, we adopted a strategy that selects the top 300 encoder features with the lowest uncertainty, ensuring a precise starting point for effective object detection. Our training methodology and hyperparameter settings closely followed those specified in RT-DETR [47]. AgeDETR was trained by using the AdamW optimizer on an NVIDIA RTX 2080 Ti GPU with a batch size of 4. The training regimen spanned 300 epochs, starting with an initial learning rate of 0.0001 and a weight decay rate of 0.0001. We trained the model with 640 × 640 pixel images to ensure comprehensive coverage of spatial details and features. We strictly followed established object detection algorithms in all experiments to achieve optimal performance and generalization capability. This approach strengthens the training process and establishes a foundation for accurate and efficient object detection in real-world applications.

### *4.3. Comparisons with Other Methods*

We validated the effectiveness of the AgeDETR model with comparative experiments on the SPARK2022 dataset and performed a detailed comparison with multiple versions of YOLOs, including YOLOv5s, YOLOv6s, YOLOv8s, and YOLOv9c. Detailed comparative data are presented in Table 1 and Figure 6. On the SPARK2022 dataset, AgeDETR demonstrated outstanding performance, with 97.7% in $mAP_{0.5}$ and 85.2% in $mAP_{0.5:0.95}$. Despite the increased computational complexity compared with YOLOv5s, YOLOv6s, and YOLOv8s, AgeDETR achieved significant improvements in $mAP_{0.5}$ and $mAP_{0.5:0.95}$.

**Table 1.** Comparison of performance of different detection models on SPARK2022 dataset.

| | **Precision (%)** | **Recall (%)** | $mAP_{0.5}$ (%) | $mAP_{0.5:0.95}$ (%) | **Parameters, M** | **GFLOPs** |
|---|---|---|---|---|---|---|
| YOLOv5s | 94.1 | 89.4 | 95.9 | 81.3 | 9.1 | 23.8 |
| YOLOv6s | 93.1 | 87.8 | 94.5 | 81.6 | 16.3 | 44.0 |
| YOLOv8s | 94.9 | 92.2 | 96.9 | 83.9 | 11.1 | 28.5 |
| YOLOv9c | 96.9 | 94.2 | 97.8 | 86.9 | 25.3 | 102.1 |
| RT-DETR | 95.5 | 91.8 | 94.6 | 81.2 | 20.1 | 58.6 |
| AgeDETR | 97.9 | 96.0 | 97.9 | 85.2 | 15.1 | 47.9 |

Compared with YOLOv9c, which achieved the same $mAP_{0.5}$ but with higher GFLOPs, AgeDETR proved to be more efficient. To substantiate our findings, we also compared AgeDETR with RT-DETR, a real-time object detection algorithm using a ResNet-18 backbone. The results show that AgeDETR achieved additional improvements of 3.3% in $mAP_{0.5}$ and 4.0% in $mAP_{0.5:0.95}$ while reducing GFLOPs by 18.7%.

These comparative analyses underscore the superiority of AgeDETR in terms of efficiency, reliability, and accuracy in object detection tasks. The enhanced performance of the model in space object detection advances research and offers promise for practical applications requiring precise and efficient detection capabilities.

To explore the performance enhancements of AgeDETR, we conducted a detailed visual analysis, as depicted in Figure 7. Through meticulous comparative analysis, it became evident that AgeDETR excels at focusing on and densely populating key feature areas. Compared with baseline models like RT-DETR and YOLO series algorithms, AgeDETR stood out by effectively reducing interference from irrelevant features and concentrating on critical feature regions. This contributed significantly to the improved accuracy and robustness of the model.

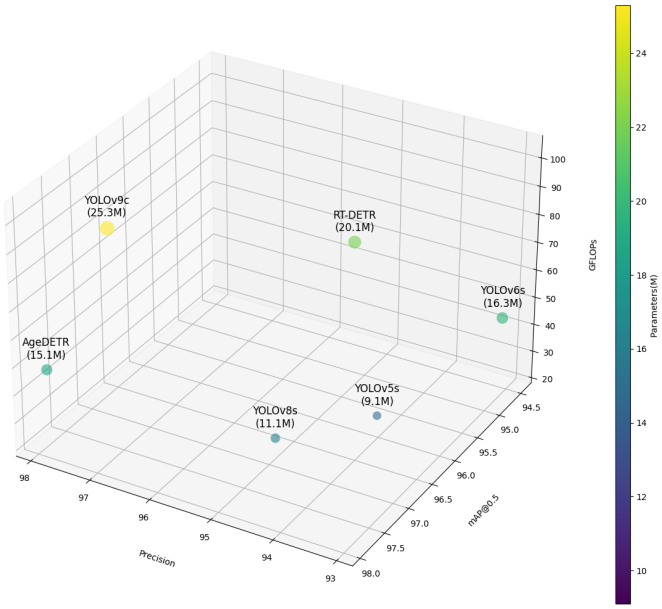
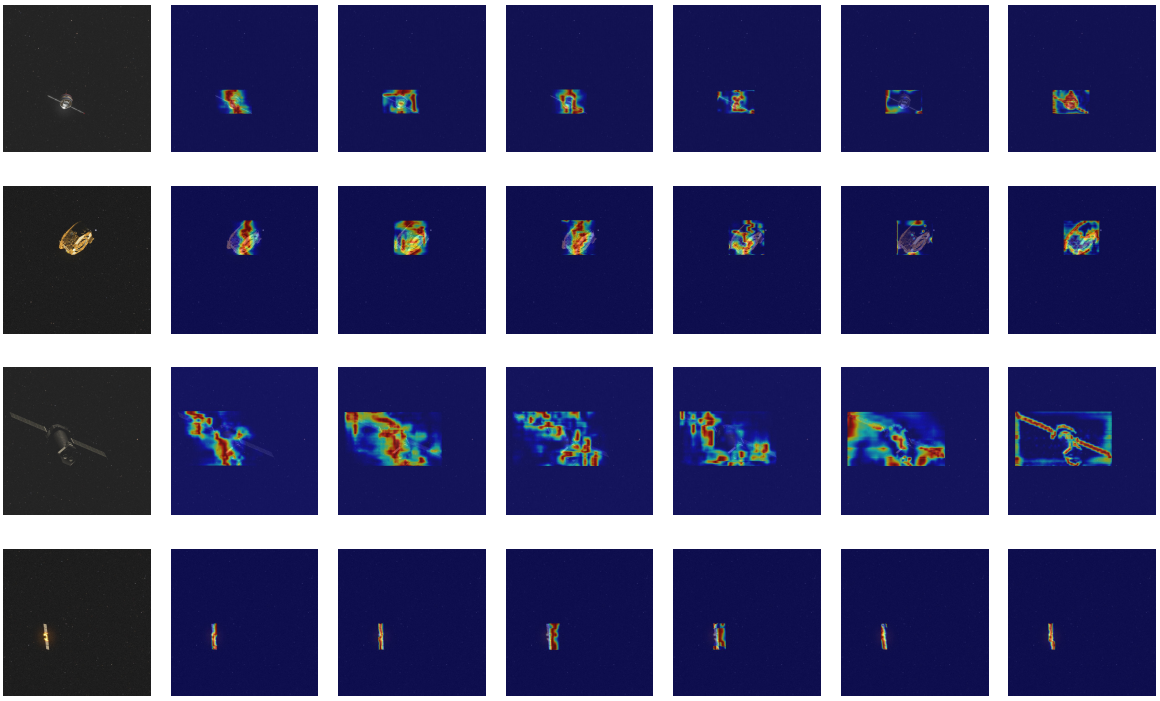



**Figure 6.** Comparison of recognition effectiveness of different detection models on SPARK2022 dataset.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**) (**g**)

**Figure 7.** Visualization of feature maps with different models. (**a**) Input image, (**b**) YOLOv5s, (**c**) YOLOv6s, (**d**) YOLOv8s, (**e**) YOLOv9c, (**f**) RT-DETR, and (**g**) AgeDETR.

Our visual findings verify the excellence of AgeDETR in feature recognition and provide actionable insights for algorithm optimization. This detailed examination enhances our understanding of operational mechanisms, laying a solid foundation for future research and model enhancements.

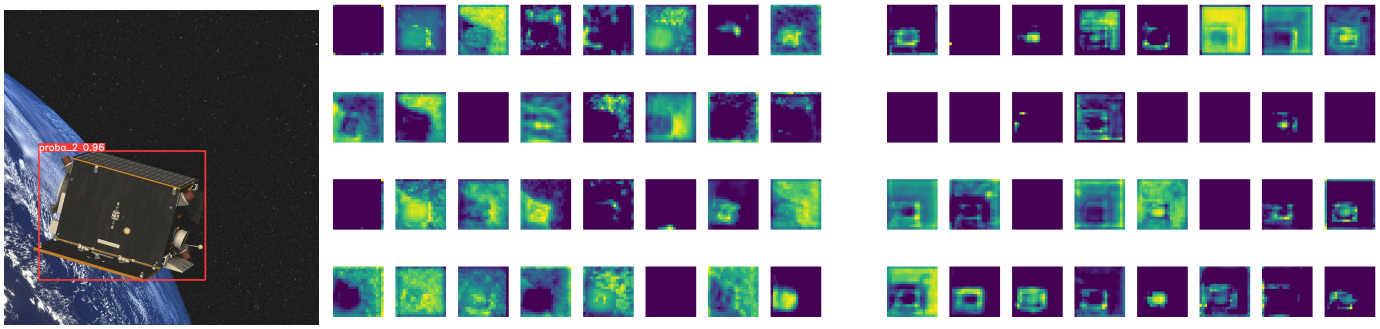
### *4.4. Ablation Studies*

In this section, we present the findings from our ablation studies, which were conducted to thoroughly evaluate the contributions of the proposed modules (EF-ResNet18, AGFE, and AGFF) within the AgeDETR framework (refer to Table 2). We initiated our

analysis by establishing a baseline model based on the AgeDETR framework, utilizing a standard ResNet18 as the backbone network while postponing the integration of the AGFE and AGFF modules. Our initial approach involved directly concatenating the multi-layer features extracted by the backbone network and feeding them into the decoder module. Figure 8a shows the input image and the label, while Figure 8b,c provide a comparative visualization of feature layer outputs from the standard ResNet18 and the EF-ResNet18 networks, respectively. The comparison reveals that the feature map of EF-ResNet18 exhibits a more pronounced response in the target region, indicating a greater focus on the target area during the feature extraction process.

**Table 2.** Comparison results of ablation experiments.

| EF-ResNet18 | AGEE | AGFF | Precision (%) | Recall (%) | $mAP_{0.5}$ (%) | $mAP_{0.5:0.95}$ (%) | Parameters, M | GFLOPs |
|---|---|---|---|---|---|---|---|---|
| × | × | × | 93.3 | 90.7 | 92.8 | 78.3 | 15.3 | 42.1 |
| ✓ | × | × | 97.2 | 95.2 | 96.6 | 83.4 | 12.3 | 36.3 |
| ✓ | ✓ | × | 97.1 | 95.8 | 96.8 | 83.7 | 13.2 | 36.4 |
| ✓ | ✓ | ✓ | 97.9 | 96.0 | 97.9 | 85.2 | 15.1 | 47.9 |




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

**Figure 8.** Visualization results of random initial weight output feature maps for different network architectures. (**a**) Input images, (**b**) ResNet18, and (**c**) EF-ResNet18.

Subsequently, to isolate the effects of each module, we systematically reintegrated them into the baseline AgeDETR model. Notably, when integrating the EF-ResNet18 module, we omitted the AGFE and AGFF modules, facilitating a straightforward concatenation of the features for input into the decoder. Throughout the experiment, we meticulously monitored and documented the training progress and prediction outputs, ensuring a comprehensive examination of the functionality.

The results indicate that the EF-ResNet18 module alone yielded a performance improvement of $+3.8$ $mAP_{0.5}$ compared with the baseline model. Furthermore, the integration of EF-ResNet18 with the AGFE module resulted in an additional performance boost of $+4.0$ $mAP_{0.5}$. Ultimately, when all modules were combined, AgeDETR exhibited exceptional overall performance, significantly enhancing the accuracy of target detection.

Figure 9 displays some of the detection results from the SPARK2022 dataset, with the upper section representing the labels of space targets and the lower section illustrating the predictions of AgeDETR. The experimental results not only validated the effectiveness of each module but also yielded profound insights crucial to further optimizing the algorithm. These findings are invaluable for deepening the understanding of the operational mechanisms of AgeDETR and will guide the refinement and advancement of future models. They serve as a cornerstone for enhancing the model's performance and capabilities in object detection tasks, establishing a solid foundation for ongoing research and development in this field.

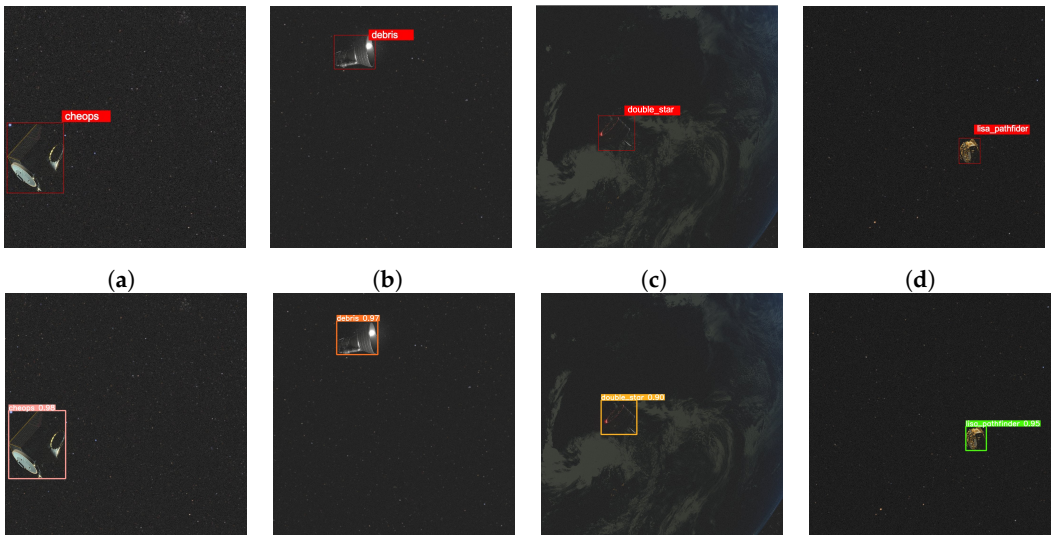



**Figure 9.** Prediction results on the SPARK2022 dataset: the upper section shows the ground truth labels for space targets, while the lower section shows the predictions by AgeDETR.

## 5. Conclusions

In this work, we propose the AgeDETR model, which significantly improves the performance of space target detection. Particularly, AgeDETR includes three critical advancements: an improved EF-ResNet18 backbone network, which boosts feature extraction capabilities and optimizes computational efficiency; the AGFE module, which enhances target feature recognition and optimizes critical information extraction; and the AGFF module, which further augments feature fusion performance. Additionally, the proposed AgeDETR eliminates the inconvenience of NMS thresholds, which facilitates practical applications. Despite challenges such as detecting weakly textured targets and reducing false positives from background noise, the effectiveness of AgeDETR is demonstrated on the SPARK2022 dataset, achieving a substantial advancement in space target detection.

**Author Contributions:** Conceptualization, X.W., B.X., C.X., and H.X.; methodology, X.W. and B.X.; software, X.W. and T.Z.; validation, X.W. and C.X.; writing—original draft, X.W.; writing—review and editing, B.X. and C.X.; visualization, X.W.; supervision, B.X. and C.X.; funding acquisition, C.X. and H.X. 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 under Grant 62401434, the China Postdoctoral Science Foundation under Grant 2023M732742, the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2024JC-YBQN-0641, Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515110504, the Postdoctoral Science Foundation of Shaanxi Province under Grant 2023BSHY-DZ797.

**Data Availability Statement:** The data used in this study are publicly available at https://doi.org/10.5281/zenodo.6599762, which can be referred to: Rathinam, Arunkumar and Gaudilliere, Vincent and Mohamed Ali, Mohamed Adel and Ortiz Del Castillo, Miguel and Pauly, Leo and Aouada, Djamila. SPARK 2022 Dataset: Spacecraft Detection and Trajectory Estimation. Zenodo.2022.

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

## References

1. Su, S.; Niu, W.; Li, Y.; Ren, C.; Peng, X.; Zheng, W.; Yang, Z. Dim and Small Space-Target Detection and Centroid Positioning Based on Motion Feature Learning. *Remote Sens.* **2023**, *15*, 2455. [CrossRef]
2. Wang, S.; Zhang, K.; Chao, L.; Chen, G.; Xia, Y.; Zhang, C. Investigating the Feasibility of Using Satellite Rainfall for the Integrated Prediction of Flood and Landslide Hazards over Shaanxi Province in Northwest China. *Remote Sens.* **2023**, *15*, 2457. [CrossRef]
3. Zhang, H.; Gao, J.; Xu, Q.; Ran, L. Applying Time-Expended Sampling to Ensemble Assimilation of Remote-Sensing Data for Short-Term Predictions of Thunderstorms. *Remote Sens.* **2023**, *15*, 2358. [CrossRef]

4. Jiang, C.; Zhao, D.; Zhang, Q.; Liu, W. A Multi-GNSS/IMU Data Fusion Algorithm Based on the Mixed Norms for Land Vehicle Applications. *Remote Sens.* **2023**, *15*, 2439. [CrossRef]
5. Saynisch, J.; Irrgang, C.; Thomas, M. On the use of satellite altimetry to detect ocean circulation's magnetic signals. *J. Geophys. Res. Ocean.* **2018**, *123*, 2305–2314. [CrossRef]
6. Kuznetsov, V.D.; Sinelnikov, V.M.; Alpert, S.N. Yakov Alpert: Sputnik-1 and the first satellite ionospheric experiment. *Adv. Space Res.* **2015**, *55*, 2833–2839. [CrossRef]
7. Buchs, R.; Florin, M.V. *Collision Risk from Space Debris: Current Status, Challenges and Response Strategies*; International Risk Governance Center: Geneva, Switzerland, 2021.
8. Johnson, N.L. Orbital debris: The growing threat to space operations. In Proceedings of the 33rd Annual Guidance and Control Conference, Breckenridge, CO, USA, 5–10 February 2010; Number AAS 10-011.
9. Tao, H.; Che, X.; Zhu, Q.; Li, X. Satellite In-Orbit Secondary Collision Risk Assessment. *Int. J. Aerosp. Eng.* **2022**, *2022*, 6358188. [CrossRef]
10. Kennewell, J.A.; Vo, B.N. An overview of space situational awareness. In Proceedings of the 16th International Conference on Information Fusion, Istanbul, Turkey, 9–12 July 2013; pp. 1029–1036.
11. McCall, G.H.; Darrah, J.H. Space Situational Awareness: Difficult, Expensive-and Necessary. *Air Space Power J.* **2014**, *28*, 6.
12. Meng, W.; Jin, T.; Zhao, X. Adaptive method of dim small object detection with heavy clutter. *Appl. Opt.* **2013**, *52*, D64–D74. [CrossRef]
13. Han, J.; Liu, S.; Qin, G.; Zhao, Q.; Zhang, H.; Li, N. A Local Contrast Method Combined With Adaptive Background Estimation for Infrared Small Target Detection. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1442–1446. [CrossRef]
14. Duk, V.; Rosenberg, L.; Ng, B.W.H. Target Detection in Sea-Clutter Using Stationary Wavelet Transforms. *IEEE Trans. Aerosp. Electron. Syst.* **2017**, *53*, 1136–1146. [CrossRef]
15. Smith, J.; Doe, J.; Zhang, W. Temporal Filtering for Enhanced Space Target Detection. *IEEE Trans. Aerosp. Electron. Syst.* **2022**, *58*, 1234–1245.
16. Liu, J.; Zhang, J.; Chen, W. Dim and Small Target Detection Based on Improved Spatio-Temporal Filtering. *IEEE Trans. Aerosp. Electron. Syst.* **2023**, *59*, 3456–3467.
17. Liu, J.; Zhang, J.; Chen, W. Infrared Moving Small Target Detection Based on Space–Time Combination in Complex Scenes. *Remote Sens.* **2023**, *15*, 5380. [CrossRef]
18. Wang, Q.; Gu, Y.; Tuia, D. Discriminative Multiple Kernel Learning for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 3912–3924. [CrossRef]
19. Wang, Q.; Wang, M.; Huang, J.; Liu, T.; Shen, T.; Gu, Y. Unsupervised Domain Adaptation for Cross-Scene Multispectral Point Cloud Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5705115. [CrossRef]
20. Wang, Q.; Wang, M.; Zhang, Z.; Song, J.; Zeng, K.; Shen, T.; Gu, Y. Multispectral Point Cloud Superpoint Segmentation. *Sci. China Technol. Sci.* **2023**, *67*, 1270–1281. [CrossRef]
21. Wang, Q.; Chi, Y.; Shen, T.; Song, J.; Zhang, Z.; Zhu, Y. Improving RGB-Infrared Object Detection by Reducing Cross-Modality Redundancy. *Remote Sens.* **2022**, *14*, 2020. [CrossRef]
22. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, L.; Wang, G.; et al. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
23. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016; Volume 1, pp. 326–366.
24. Xue, D.; Sun, J.; Hu, Y.; Zheng, Y.; Zhu, Y.; Zhang, Y. Dim small target detection based on convolutinal neural network in star image. *Multimed. Tools Appl.* **2020**, *79*, 4681–4698. [CrossRef]
25. Xiang, Y.; Xi, J.; Cong, M.; Yang, Y.; Ren, C.; Han, L. Space debris detection with fast grid-based learning. In Proceedings of the 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI), Chongqing City, China, 28–30 November 2020; pp. 205–209. [CrossRef]
26. Xi, J.; Xiang, Y.; Ersoy, O.K.; Cong, M.; Wei, X.; Gu, J. Space Debris Detection Using Feature Learning of Candidate Regions in Optical Image Sequences. *IEEE Access* **2020**, *8*, 150864–150877. [CrossRef]
27. Redmon, S.; Divvala, R.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
28. Redmon, J.; Farhadi, A. Yolo9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
29. Farhadi, A.; Hejrati, B.; Ravanbakhsh, M.; Bagheri, Y.; Ghodrati, A.; Davoodi, S.; Sedghi, M. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
30. Jocher, G.; Ultralytics. YOLOv5. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 5 September 2023).
31. 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.
32. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Yolov7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In Proceedings of the IEEE/Cvf Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022.

33. Varghese, R.; Sambath, M. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness. In Proceedings of the 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 18–19 April 2024; pp. 1–6. [CrossRef]
34. Wang, C.-Y.; Yeh, I.-H.; Liao, H.-Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. *arXiv* **2024**, arXiv:2402.13616.
35. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
36. Wang, Y.; Zhang, X.; Yang, T.; Sun, J. Anchor DETR: Query design for transformer-based detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022.
37. Cao, X.; Yuan, P.; Feng, B.; Niu, K. DQ-DETR: Dual Query Detection Transformer for Phrase Extraction and Grounding. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.
38. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional DETR for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
39. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. Dn-DETR: Accelerate DETR training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 13619–13627.
40. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *arXiv* **2022**, arXiv:2203.03605.
41. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. *arXiv* **2020**, arXiv:2010.04159.
42. Gao, P.; Zheng, M.; Wang, X.; Dai, J.; Li, H. Fast Convergence of DETR with Spatially Modulated Co-Attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
43. Sun, Z.; Cao, S.; Yang, Y.; Kitani, K. Rethinking transformer-based set prediction for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seattle, WA, USA, 13–19 June 2020.
44. Dai, X.; Chen, Y.; Yang, J.; Zhang, P.; Yuan, L.; Zhang, L. Dynamic DETR: End-to-end object detection with dynamic attention. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 2968–2977.
45. Cao, X.; Yuan, P.; Feng, B.; Niu, K. Cf-DETR: Coarse-to-fine transformers for end-to-end object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 22 February–1 March 2022.
46. JustIC03. MFDS-DETR: Multi-level Feature Fusion with Deformable Self-Attention for White Blood Cell Detection. *arXiv* **2022**, arXiv:2212.11659.
47. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 16965–16974.
48. Pauly, L.; Jamrozik, M.L.; Del Castillo, M.O.; Borgue, O.; Singh, I.P.; Makhdoomi, M.R.; Christidi-Loumpasefski, O.O.; Gaudilliere, V.; Martinez, C.; Rathinam, A.; et al. Lessons from a Space Lab—An Image Acquisition Perspective. *Int. J. Aerosp. Eng.* **2023**, *2023*, 9944614. [CrossRef]
49. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.; Chan, S. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 17–24 June 2023; pp. 12021–12031. [CrossRef]
50. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes, Greece, 4–10 June 2023; pp. 1–5. [CrossRef]
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
52. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10687–10698.
53. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; Volume 1807.
54. Yao, Z.; Ai, J.; Li, B.; Zhang, C. Efficient DETR: Improving End-to-End Object Detector with Dense Prior. *arXiv* **2021**, arXiv:2104.01318.
55. Li, F.; Zeng, A.; Liu, S.; Zhang, H.; Li, H.; Zhang, L.; Ni, L.M. Lite DETR: An Interleaved Multi-Scale Encoder for Efficient DETR. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023.
56. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. Dynamic Anchor Boxes are Better Queries for DETR. In Proceedings of the International Conference on Learning Representations (ICLR), Online, 25 April 2022.
57. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141. [CrossRef]
58. Gao, Z.L.; Xie, J.T.; Wang, Q.L.; Li, P.H. Global Second-Order Pooling Convolutional Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3019–3028.

59. Lee, H.; Kim, H.E.; Nam, H. SRM: A Style-Based Recalibration Module for Convolutional Neural Networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Long Beach, CA, USA, 15–20 June 2019; pp. 1854–1862.
60. Wang, Q.L.; Wu, B.G.; Zhu, P.F.; Li, P.H.; Zuo, W.M.; Hu, Q.H. ECA-net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11531–11539.
61. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 13–18 December 2014; Volume 2, pp. 2204–2212.
62. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial Transformer Networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; Volume 28, pp. 37–45.
63. Wang, F.; Jiang, M.Q.; Qian, C.; Yang, S.; Li, C.; Zhang, H.G.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6450–6458.
64. Park, A.; OtherAuthor, A. Bottleneck Attention Module for Efficient Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7459–7468.
65. Liu, S.; Qi, X.; Qin, H.; Shi, J.; Jia, J. CBNet: A Novel Composite Backbone Network Architecture for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10512–10521.
66. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
67. 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 16x16 words: Transformers for image recognition at scale. In Proceedings of the 9th International Conference on Learning Representations, Online, 3–7 May 2021.
68. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
69. Yin, M.; Yao, Z.; Cao, Y.; Li, X.; Zhang, Z.; Lin, S.; Hu, H. Disentangled non-local neural networks. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2020; Volume 12360, pp. 191–207.
70. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.