remote sensing

MDPI

Article

Enhancing Object Detection in Remote Sensing: A Hybrid YOLOv7 and Transformer Approach with Automatic Model Selection

Mahmoud Ahmed [1,*], Naser El-Sheimy [2], Henry Leung [1] and Adel Moussa [2,3]

[1] Department of Electrical and Software Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada; leungh@ucalgary.ca
[2] Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada; elsheimy@ucalgary.ca (N.E.-S.); amelsaye@ucalgary.ca (A.M.)
[3] Department of Electrical and Computer Engineering, Port-Said University, Port-Said 42523, Egypt
* Correspondence: mahmoud.ahmed2@ucalgary.ca

Abstract: In the remote sensing field, object detection holds immense value for applications such as land use classification, disaster monitoring, and infrastructure planning, where accurate and efficient identification of objects within images is essential for informed decision making. However, achieving object localization with high precision can be challenging even if minor errors exist at the pixel level, which can significantly impact the ground distance measurements. To address this critical challenge, our research introduces an innovative hybrid approach that combines the capabilities of the You Only Look Once version 7 (YOLOv7) and DEtection TRansformer (DETR) algorithms. By bridging the gap between local receptive field and global context, our approach not only enhances overall object detection accuracy, but also promotes precise object localization, a key requirement in the field of remote sensing. Furthermore, a key advantage of our approach is the introduction of an automatic selection module which serves as an intelligent decision-making component. This module optimizes the selection process between YOLOv7 and DETR, and further improves object detection accuracy. Finally, we validate the improved performance of our new hybrid approach through empirical experimentation, and thus confirm its contribution to the field of target recognition and detection in remote sensing images.

Keywords: object detection; detection transformer; YOLOv7; multimodalities

Citation: Ahmed, M.; El-Sheimy, N.; Leung, H.; Moussa, A. Enhancing Object Detection in Remote Sensing: A Hybrid YOLOv7 and Transformer Approach with Automatic Model Selection. *Remote Sens.* **2024**, *16*, 51. https://doi.org/10.3390/rs16010051

Academic Editors: Pedram Ghamisi, Ce Zhang, Danfeng Hong and Qiqi Zhu

Received: 11 November 2023
Revised: 19 December 2023
Accepted: 20 December 2023
Published: 21 December 2023

1. Introduction

Detecting objects in remote sensing images presents a set of intricate challenges that demand innovative solutions. Remote sensing data typically originate from diverse sensor modalities, such as optical, synthetic aperture radar (SAR), or multispectral sensors. Each modality possesses unique characteristics and intricacies, thus necessitating adaptable detection methods to select the optimal one for the situation at hand [1]. Additionally, remote sensing images exhibit high variability in environmental conditions, such as changes in lighting and weather, which poses challenges in the development of robust models for all scenarios. Not only that, the vast geographical areas covered by such images also require algorithms capable of efficiently processing large datasets.

The fusion of data from multiple sensors introduces complexities due to spatial, spectral, and temporal resolution variations. Geo-referencing remote sensing images poses difficulties related to geometric distortions, registration, and non-linear distortions caused by the Earth's curvature [1,2]. Moreover, the presence of limited annotated training data (which are often costly to acquire) impedes accurate model development, especially in the context of rare and specific object detection. Detecting small objects and resolving class imbalance issues further complicates the challenge, given that the number of objects of

interest is typically much smaller than the number of background objects [3]. Finally, it is noted that the multifaceted nature of object detection in remote sensing images is influenced by factors such as adaptability to diverse object types, considerations of real-time processing, semantic context understanding, as well as privacy concerns when capturing sensitive information [4]. Therefore, in order to overcome these challenges and thus drive innovation in this critical field, the integration of advanced machine learning, domain expertise, and a deep understanding of the remote sensing data intricacies are absolutely necessary [5].

Deep learning, particularly convolutional neural networks (CNNs), has had a significant impact on many aspects of computer vision, including tasks like object recognition, detection, and segmentation, while it has also sparked inspiration in fields like remote sensing [1,3]. These networks typically take RGB images as input and apply a series of operations, including convolution, local normalization, and pooling [2]. Furthermore, CNNs rely heavily on extensive training data, and the resulting pre-trained models are then used as versatile feature extractors for various downstream applications [1].

One of the fundamental components of CNNs is the convolution operation, which is essential for capturing local interactions like contour and edge information within input images [4]. CNNs incorporate inductive biases such as spatial connectivity and translation equivariance, contributing to the construction of robust and efficient architectures. However, a limitation of CNNs is their local receptive field, which restricts their ability to model long-range dependencies in an image, such as relationships between distant parts [5]. Furthermore, convolutions are content-independent because they use fixed filter weights that are applied uniformly to all inputs, regardless of their characteristics [6].

Recently, vision transformers (ViTs) have emerged as a compelling alternative in the field of computer vision. They are based on the self-attention mechanism, which effectively captures global interactions by learning how elements in a sequence relate to one another [7]. New studies have also highlighted that ViTs excel in modeling content-dependent and long-range interactions, while dynamically adapting their receptive fields to mitigate data-related issues and thus, ultimately, learning highly effective feature representations [8,9]. Overall, ViTs and their various adaptations have proven successful in a wide range of computer vision tasks, including image classification, object detection, and image segmentation [10].

CNN-based methods typically struggle with capturing global features. In response to this limitation, the DETR model has been developed, and its effectiveness in capturing long-range relationships in remote sensing images (RSIs) has been established [11]. Many of these transformer-based approaches draw inspiration from the YOLO object detection paradigm, known for its exceptional speed, as it streamlines the process by directly predicting bounding-box coordinates and categories. As a result, this eliminates the need for region proposal search, and thus leads to significantly faster inference times [12]. Moreover, YOLOv7, a variant of the YOLO algorithm, has demonstrated improved small object recognition and robustness to different backgrounds [13]. In Ref. [14], a comparative analysis noted that YOLO achieves a commendable equilibrium between detection accuracy and computational efficiency in contrast to other CNN-based detectors within the context of RSIs.

As it was mentioned earlier, object detection plays a pivotal role in various remote sensing applications, ranging from land use classification and disaster monitoring to infrastructure planning. In this domain, achieving high accuracy in object localization is of paramount importance, as even minor errors at the pixel level can lead to significant discrepancies in ground distance measurements [15]. With the growing demand for precise and reliable object detection in remote sensing, novel approaches combining the strengths of different detection methods have emerged. In this research work, we present a hybrid approach that leverages the strengths of both YOLOv7 and the DETR, two state-of-the-art object detection frameworks [16].

1.1. Related Work

1.1.1. General Object Detection

The field of object detection has seen significant advancements with the rise of deep learning. Object detectors can be broadly categorized into two types: those with a region-of-interest proposal step (two-stage) [17] and those without (one-stage) [18]. Notably, the YOLO algorithm family, has gained attention due to their efficiency and simplicity in the one-stage design, incorporating advanced technologies like the spatial pyramid pooling (SPP) module in YOLOv3 [19] and the Mish activation function in YOLOv4 [20].

1.1.2. Vision Transformer (ViT)

Transformers, originally developed for natural language processing (NLP) tasks [21], have shown remarkable success in tasks such as machine translation, question answering, text classification, and document summarization [22]. This success is attributed to the transformer's ability to capture complex dependencies through a self-attention mechanism. Notably, the ViT [7] extended the applicability of transformers to images by treating images as sequences of patches, demonstrating a performance competitive with that of the CNN method in image recognition tasks. Additionally, the DETR was the first successful attempt to employ transformers for object detection, combining a transformer encoder and decoder with a standard CNN model and utilizing a set-matching loss function [11].

1.1.3. Hybrid Approaches

CNN methods struggle with capturing global contextual information due to the inherent locality of the convolution operation. In contrast, transformers excel in globally attending to interdependencies among image feature patches through multi-head self-attention, and thus manage to preserve ample spatial details that are essential for effective object detection [23].

Hybrid approaches in remote sensing, which combine the capabilities of transformers and CNNs, have gained increasing attention for their potential to revolutionize object detection and image analysis [21]. These approaches offer the promise of enhanced accuracy and versatility, yet they are accompanied by several notable challenges. The intricate self-attention mechanisms of transformers, while powerful, increase model complexity and computational demands [24]. Consequently, this extends training times and raises resource requirements, making it imperative to explore efficient training strategies and model architectures that are less resource-intensive [25]. One of the fundamental challenges is the need for substantial and diverse datasets, which are often scarce in remote sensing due to the high costs and effort involved in data collection and annotation [26]. Moreover, the integration of large-scale hybrid models into existing workflows demands compatibility with geospatial tools, data formats, and standards [11], necessitating the development of bridges between computer vision and geospatial domains. On another note, coordinating the training of both transformers and CNNs within hybrid models is a complex task, requiring specialized techniques to ensure effective knowledge transfer and regularization strategies to mitigate overfitting during fine-tuning for domain-specific tasks [27]. Finally, achieving interpretability and explain ability in hybrid models is crucial for transparent decision making [28], but it remains an ongoing challenge [29].

Despite the aforementioned obstacles, the field of remote sensing actively embraces hybrid models for their potential to address long-standing problems and push the boundaries of what is achievable in many applications. In addition, the use of complex and opaque black-box models, like deep neural networks and ensembles, in machine learning, raises challenges in understanding their decision-making processes. Additionally, even though there are the so-called explainers, who aim to elucidate these processes, they have limitations such as imperfect explanation fidelity and ambiguity in their explanations [28,30].

Generally, the literature contains several works that integrate transformers and CNNs. One approach introduces a Swin transformer-based backbone to enhance local perception as well as to leverage the strengths of both methods for improved local feature extraction [31].

Another study combined multi-scale global and local information from transformers and CNNs using an adaptive feature fusion network, in order to capture a comprehensive range of contextual information [32]. Ref. [16] presented a Siamese U-shaped network structure based on the Swin transformer, incorporating encoder, fusion, and decoder modules to facilitate more effective information integration. Ref. [33] proposed a framework that integrates transformer and UNet architectures to capture enriched contextualized features, up-sample them, and then fuse them with multi-scale features in order to eventually generate global–local features for enhanced image understanding. Ref. [34] suggested a hybrid CNN–transformers framework design for crop segmentation that effectively merges local details and global contextual information to improve segmentation accuracy. Ref. [35] developed an approach for remote sensing image caption generation, which involved the adaptation of transformers with residual connections, dropout, and adaptive feature fusion in order to enable more precise and context-aware captions.

1.2. Contributions

The novelty and contributions of the proposed hybrid approach for object detection in remote sensing applications can be summarized as follows:

- Integration of YOLOv7 and Detection Transformer We seamlessly integrate YOLOv7 and detection transformers, leveraging their complementary features. YOLOv7 excels in local object detection, while detection transformers provide global context, enhancing object detection by combining local precision and global awareness.
- Automatic Selection Module Our automatic model selection module, trained on the mean average precision $mAP_{0.5:0.95}$-based detection accuracy scores, serves as a decision-making component that optimizes the choice between YOLOv7 and detection transformers. It achieves higher localization accuracy by selecting the most suitable model for each scenario, a vital innovation in remote sensing object detection. This module serves as a decision-making component that optimizes the choice between YOLOv7 and detection transformer.
- Improved Object Localization in Remote Sensing The main achieved benefit as per the results is reaching higher object localization accuracy, as measured by the $mAP_{0.5:0.95}$ metric values of the proposed approach while the achieved $mAP_{0.5}$ values are comparable to the individual benchmark models. This addresses the critical requirement in remote sensing, where slight pixel-level errors can lead to significant ground distance discrepancies. Our hybrid approach consistently improves $mAP_{0.5:0.95}$ scores, benefiting object detection, land use classification, disaster monitoring, and infrastructure planning.
- Benchmark-Beating Results Our hybrid approach outperforms individual models, establishing itself as an innovative solution in remote sensing object detection. The enhancements in $mAP_{0.5:0.95}$ scores underscore the practical relevance and applicability of this research in real-world remote sensing tasks, contributing significantly to advancing geospatial information extraction.

2. Materials and Methods

While current methods such as the YOLOv7 adopt a local receptive field during prediction, others like the DETR employ the attention mechanism to perform a global context-informed manner of prediction. In this work, a hybrid approach is proposed that combines the merits of the two approaches by directing the prediction towards the most suitable model for each input image. Therefore, the images that could benefit from the local perceptive field will be processed by YOLOv7, and the images that need global context will be processed by DETR. As for the decision between the two strategies for each image, an automatic selection algorithm is brought forward that depends on the detection performance of the individual images and aims to create a selection model based on the training data without un-explained assumptions. Hence, a straightforward, informed, and

valid decision can be made. Below, a detailed description of our proposed algorithm will be provided.

The proposed workflow for object detection commences with an automatic selection (AS) module. During the training phase, this module was exposed to a diverse training dataset that includes images analyzed by both YOLOv7 and DETR methods, with $mAP_{0.5:0.95}$ scores serving as the key metric. Therefore, by training the module based on this dataset, it acquires the expertise to autonomously analyze the characteristics of new test images and select the strategy that maximizes accuracy. This adaptive decision-making process significantly enhances the detection performance, an essential factor for promoting precision in remote sensing.

Our new method also integrates a convolutional transformer detection block (CTDB), where the DETR captures the global context, is pretrained on contextual representations and fine-tuned for the specific detection task, and effectively utilizes dilated backbones to improve small object detection [11]. Moreover, the DETR passes the input embeddings to the transformer's encoder–decoder, which generates output embeddings. Finally, the DETR passes each output embedding to a classifier feed-forward network (FFN) and a bounding box FFN for producing the final predictions, as shown in Figure 1 [11]. Meanwhile, YOLOv7 incorporates extended–efficient layer aggregation networks (E-ELAN) enhancements as well as novel model scaling to ensure accurate local object detection [36]. Together, these components form a comprehensive and intelligent object detection workflow tailored to the unique demands of remote sensing applications. Up next, our hybrid structure for object detection will be presented, where Figure 2 illustrates the architecture of our method, Section 2.1. highlights the decision-making process via the automatic selection module, and Section 2.2. shows the CTDB.

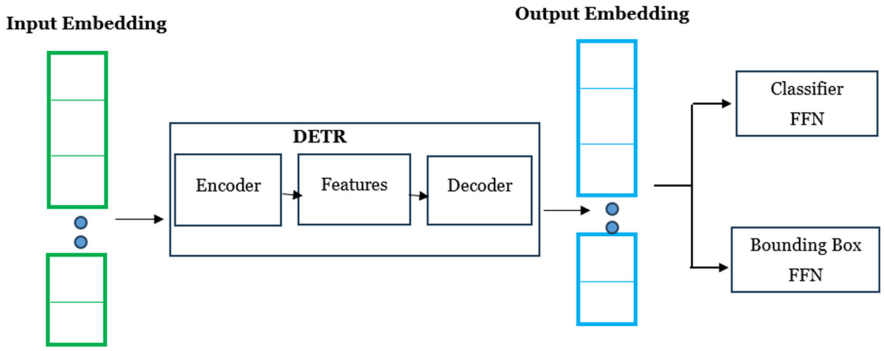

Figure 1. DETR pipeline overview.

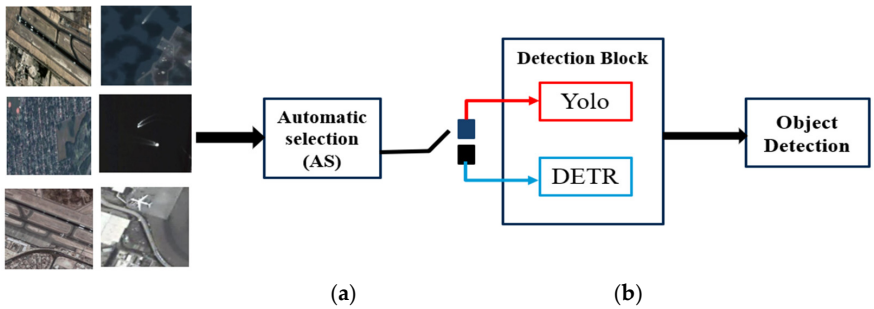

Figure 2. Hybrid structure for object detection, where (**a**) is the automatic selection module and (**b**) is the convolutional transformer detection block.

2.1. Decision-Making Process

In our methodology, the AS module is primarily reliant on predictions generated by the detection models, without any need for modifying the detection models or acquiring additional information from them. This minimal information requirement provides enhanced flexibility when establishing hybrid approaches between different models. Moreover, this

approach streamlines the collaboration process and ensures the accuracy of information within the model.

The AS module, a pivotal component within the hybrid object detection structure illustrated in Figure 2a, assumes a critical role in determining the application of either YOLOv7 or DETR to a given test image. As for its configuration, a comprehensive evaluation was conducted, where various modules such as the EfficientNetB0, InceptionV3, ResNet50, ResNet101, and Darknet53 were considered, with each bringing its own unique characteristics to the forefront.

EfficientNetB0, known for its efficiency, balances computational resources and accuracy [37]. InceptionV3, characterized by its inception modules, demonstrates versatility in capturing complex features [38]. ResNet50, a residual network architecture, showcases improved training ease and accuracy, especially when in the context of deep networks [39]. ResNet101, an extension of ResNet50, offers further depth and feature representation capabilities [9]. Darknet53, known for its role in YOLO models, emphasizes efficient object detection through its unique architecture [40].

The aforementioned comprehensive empirical approach highlights the significance of fine-tuning the model selection process based on experimental results, while also considering the nuanced strengths of different model architectures in the pursuit of optimal performance for the object detection task at hand [11,41]. In fact, this evaluation identified ResNet101 as the optimal choice for the AS module, since it outperformed its counterparts in terms of both computational efficiency and accuracy.

Hereinafter, we provide a step-by-step description of the decision-making process, implemented by the AS module, in the context of the proposed object detection algorithm, with the purpose of determining the most appropriate detection model for an input image.

2.1.1. Automatic Selection Module (AS)

The first step in the decision-making process is the operation of the AS module, denoted by the input image represented as (x). AS is responsible for evaluating the characteristics of the input image and utilizing its training to make predictions about the performance of the available detection models. Moreover, in the context of our algorithm, the $mAP_{0.5:0.95}$ quantity has been chosen for evaluation of the automatic selection model training quality. The reason behind our choice is because it is a comprehensive measure of detection accuracy across different intersection over union (IoU) thresholds, which means that the $mAP_{0.5:0.95}$ accounts for recall, precision, and localization of the prediction. Therefore, in this way, the module learns from these training data how the performance of each model varies on different images, while also considering factors such as image content, complexity, and object distribution.

During the training, the automatic selection modules are exposed to a diverse set of training data. This dataset includes various images for which object detection has been performed using both YOLOv7 and DETR models. In fact, these images are used as input to our proposed AS model. Subsequently, for each input image (x) in the training dataset, the automatic selection model predicts the expected mean average precision values $mAP_{0.5:0.95\ YOLOv7}$ and $mAP_{0.5:0.95\ DETR}$ [42].

For a given input image (x), the automatic selection model predicts the expected mean average precision values $mAP_{0.5:0.95\ YOLOv7}$ and $mAP_{0.5:0.95\ DETR}$ [42]. Specifically, the $mAP_{0.5:0.95}$ of YOLOV7 for the image (x) can be calculated as follows:

$$mAP_{0.5:0.95\ YOLOv7} = \frac{1}{10}\sum_{i=1}^{10} AP_{0.5+0.05\cdot i}\ , \tag{1}$$

while the $mAP_{0.5:0.95}$ of DETR for the input image (x) is given by the following equation:

$$mAP_{0.5:0.95\ DETR} = \frac{1}{10}\sum_{i=1}^{10} AP_{0.5+0.05\cdot i} \tag{2}$$

2.1.2. Decision-Making Phase

The decision-making phase is based on an informed selection based on the predicted $mAP_{0.5:0.95}$ values, and it is essential to the autonomous operation of the system. In the following, the mathematical expression of the criterion for the selection of the appropriate detection model is denoted as M, and it is defined as follows:

$$M = \begin{cases} \text{YOLOv7} & \text{if } mAP_{0.5:0.95\ YOLOv7} > mAP_{0.5:0.95\ DETR} \\ \text{DETR} & \text{otherwise} \end{cases} \quad (3)$$

Based on Equation (3), it is evident that if the predicted $mAP_{0.5:0.95}$ value for the YOLOv7 method is larger than the one for DETR, then the input image is directed to YOLOv7. In any other case, it is guided towards the DETR method.

This algorithm is a structured and automated approach that enhances detection accuracy by adapting to the unique characteristics of the image at hand. This innovative automatic selection module leverages performance metrics from both the YOLOv7 and DETR methods to ensure an autonomous choice of the most appropriate detection strategy. This adaptability contributes to the improvement of detection performance, particularly in remote sensing scenarios, where precision is paramount due to the direct correlation between pixel-level errors in the images and ground-level distances. Essentially, this intelligent module acts as a mechanism for precise and reliable object detection by making the optimal processing strategy selection for each specific case. The automatic selection module is a key innovation that leverages performance metrics from both YOLOv7 and DETR to autonomously guide the selection of the most appropriate detection model for each image. Its capacity to adapt and optimize model choice based on image content significantly enhances object localization accuracy, making it an asset in remote sensing and object detection.

2.2. Convolutional Transformer Detection Block (CTDB)

The CTDB consists of two object detection modules, the DETR and the YOLOv7, while its general architecture is depicted in Figure 2b.

2.2.1. DETR-DC5-R101 Model Architecture

The DETR-DC5-R101 model represents a significant extension of the DETR architecture, introducing a DC5-R101 backbone that enhances feature extraction and contextual awareness. This integration is a strategic choice that addresses the common challenges faced by traditional object detection methods. These challenges include difficulties adapting to variations in object scale, complex object shapes, and cluttered scenes with overlapping objects. On another note, the DC5-R101 backbone offers solutions by enhancing adaptability, feature extraction, and contextual awareness. As a result, this ensures that the DETR model with the DC5-R101 backbone remains at the forefront of object detection capabilities [11,43].

The model's architectural flow initiates with the input image (x), complemented by the incorporation of positional encodings $PE(\mathrm{x})$ to precisely capture object positions. In addition, positional encodings are introduced to the input feature maps, allowing the model to understand spatial relationships. These encodings can be defined as follows for a 2D position (i, j) in the feature map, and more specifically for positional encodings [44]:

$$\begin{aligned} PE(i,j) = [&\sin(i/10000\hat{}(2\star d/emb\_dim)), \cos(i/10000\hat{}(2\star d/emb\_dim)) \\ &\sin(j/10000\hat{}(2\star d/emb\_dim)), \cos(j/10000\hat{}(2\star d/emb\_dim))] \,, \end{aligned} \quad (4)$$

where i and j represent the row and column indices in the feature map, while $2\star d$ and emb_dim refer to the number of dimensions in the positional encoding.

The DC5-R101 backbone, a principal component of this architecture, excels in its ability to adaptively adjust the receptive field of feature maps, a fundamental requirement for robust object detection. This adjustment is achieved through the introduction of dilations

and stride modifications, leading to an enhanced feature map output. This operation can be represented using the following equation [11]:

$$\mathcal{H}_{backbone} = DC5{-}R101(PE(\mathrm{x})) \tag{5}$$

where $\mathcal{H}_{backbone}$ represents the output of the DC5-R101 backbone, and $PE(\mathrm{x})$ is the input image with positional encodings. However, this architectural enhancement comes at the trade-off of increased computational cost due to intensified self-attention mechanisms in the encoder. As for the DC5 operation in this context, it is mathematically represented as follows:

$$DC5(Y, d) = Conv(Y, 5 \times 5,\ dilation = d) \tag{6}$$

Based on Equation (6), DETR leverages a CNN with a kernel (5×5) and a specified dilation rate (d) to enhance adaptability for fine-grained detail capture. The kernel (5×5) is chosen for its effectiveness in capturing intricate patterns and spatial relationships. As for the dilation rate, it introduces adaptability to the convolutional operation, influencing receptive field construction. This approach enables the network to process information at multiple scales, crucial for tasks requiring detailed analysis.

Furthermore, the DETR architecture employs a transformer encoder that incorporates self-attention and feed-forward layers. Within this encoder, the self-attention mechanism calculates attention scores for each object query–key pair (Q_i, K_i) through the formula below [21]:

$$Attention(Q_i, K_i) = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \cdot V \tag{7}$$

where d_k denotes the dimension of keys and values, and V represents the values, while the transformer encoder output is represented as $\mathcal{H}_{transformer}$.

The model makes predictions for each object query using feedforward layers. Specifically, class predictions (P_{cls}) and bounding box predictions (P_{box}) are generated using the following equations [21]:

$$P_{\text{cls}} = \text{Softmax}\left(W_c \cdot \mathcal{H}_{transformer}\right) \tag{8}$$

$$P_{\text{box}} = W_{box} \cdot \mathcal{H}_{transformer} \tag{9}$$

where W_c and W_{box} are the learned weight matrices.

The model is trained using a combination of loss functions $(Loss_{total})$, including a classification loss $(Loss_{\text{cls}})$ and a localization loss $(Loss_{\text{box}})$ [11]:

$$Loss_{total} = Loss_{\text{cls}} + Loss_{\text{box}} \tag{10}$$

In object detection models, the terms "classification loss" and "localization loss" refer to two essential components of the overall loss function, which guides the training of the model. $Loss_{\text{cls}}$ is crucial in object detection since it measures the accuracy of predicting object class labels by quantifying the difference between predicted and ground truth labels. Furthermore, this quantity is calculated with a loss function like cross-entropy; it guides the model to assign high probabilities to correct class labels, thereby ensuring accurate identification of object categories. As for the $Loss_{\text{box}}$, it assesses the accuracy of predicting bounding box coordinates for objects in an image. In fact, it gauges how well the predicted bounding boxes align with the actual object positions. Moreover, it is typically calculated with smooth L_1 loss or a similar regression function and encourages the model to precisely locate objects in the image by minimizing disparities between predicted and ground truth bounding boxes. Finally, the $Loss_{total}$ combines the $Loss_{\text{cls}}$ and $Loss_{\text{box}}$ to create a joint optimization goal for training. Basically, the DETR architecture aims to simultaneously minimize both components and thus achieve a balance between accurate object classification and precise localization within the image.

2.2.2. Convolutional Neural Network YOLOv7

YOLOv7, the latest iteration of the YOLO series, represents a significant advancement in both object detection speed and accuracy compared to its predecessors. Its key improvement lies in the overall architecture, where the concept of E-ELAN has been introduced.

This method features a series of convolutional layers, followed by detection layers that are responsible for predicting bounding boxes and object class probabilities. In addition, it utilizes techniques like expansion, shuffling and cardinality merging to enhance the network's learning capacity continuously. Most importantly, this augmentation does not disrupt the original gradient flow, ensuring stable training. In fact, E-ELAN has been found to be particularly effective in guiding different sets of computational blocks in order to acquire diverse features, as shown in Figure 3a [45]. Finally, the YOLOv7 introduces a novel approach to model scaling, which aims to preserve the model's inherent characteristics from its initial design, while also maintaining the optimal structure, as shown in Figure 3b. Overall, these innovations represent a significant stride forward in the YOLO series framework in terms of network optimization and enhancement.

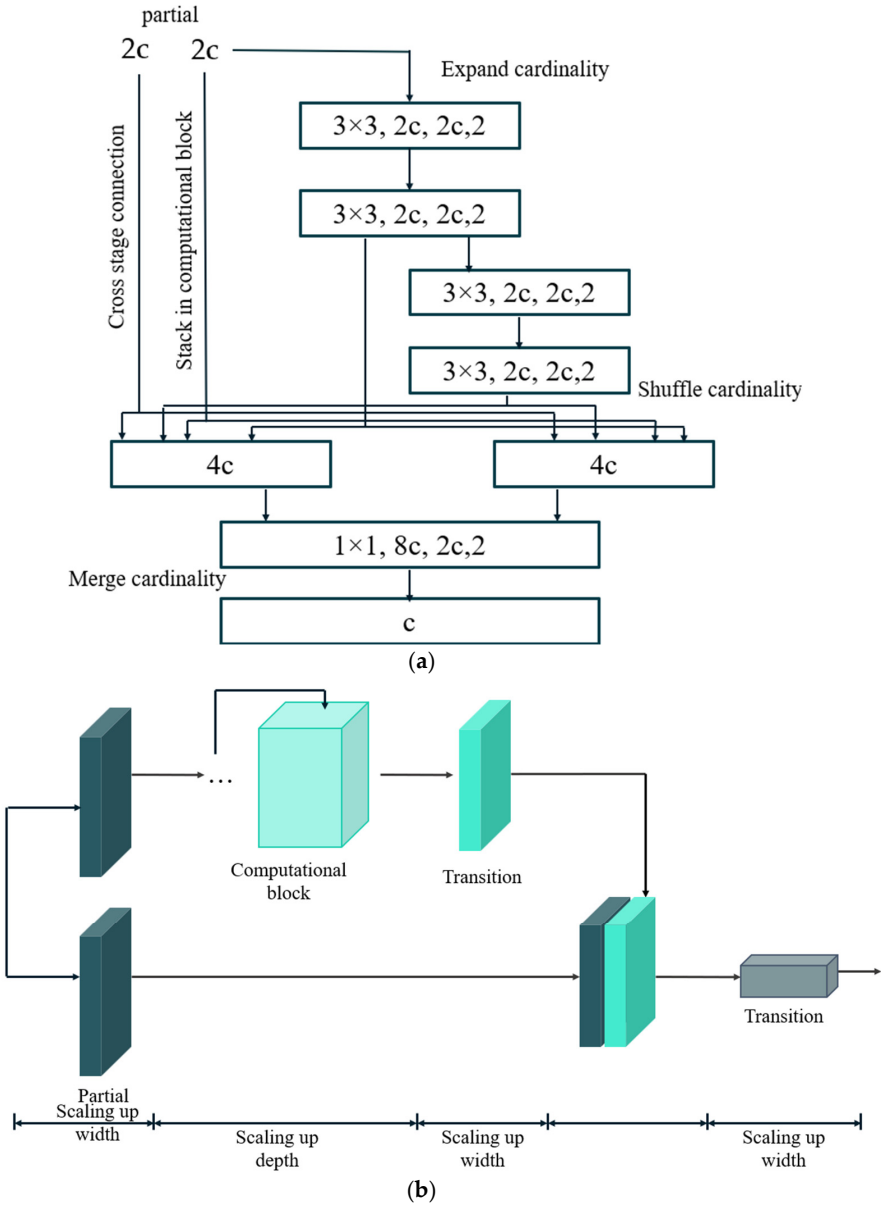

Figure 3. The key components of YOLOV7, where (**a**) are the E-ELAN structures, and (**b**) is the model scaling for a concatenation-based model.

2.3. Materials

In this study, the specialized VIdeo Satellite Objects (VISO) dataset was utilized for the precise detection of objects within satellite images [46]. The VISO dataset comprises an extensive collection designed for the task of detecting and tracking moving objects within satellite videos. This dataset includes 47 satellite video sequences, all of which were captured by Jilin-1 satellite platforms. Each image in this dataset boasts a high resolution, measuring 12,000 pixels in width and 5000 pixels in height. Furthermore, this extensive collection not only provides a diverse range of objects but also encompasses various sizes and scales, enriching the dataset with real-world complexities. In our work here, we focus on the detection of three common classes: airplanes, ships, and trains.

The training process utilized a computer with a 12th Gen Intel(R) Core (TM) i7-12700H processor running at 2.70 GHz, 16.0 GB of installed RAM (15.7 GB usable), and a GeForce RTX 3070 graphics card. Furthermore, the deep learning framework employed for this work was Pytorch.

3. Results

As it was mentioned earlier, we integrated different deep learning neural networks into the AS module in order to assess their object detection performance in terms of accuracy and determine the optimal one. Specifically, the modules we considered were the EfficientNetB0, InceptionV3, ResNet50, ResNet101, and Darknet53. The results of this evaluation are summarized in Table 1. Based on that, ResNet101 demonstrated superior accuracy compared to others, and thus it is designated as the one capable of producing research outcomes with optimal robustness and credibility.

Table 1. Accuracy of the automatic selection module using different convolutional neural networks for different objects.

Methods	Accuracy of Selection According to Objects		
	Plane	**Ship**	**Train**
Resnet101	85%	83.7%	88.57%
Efficientnetb0	84.45%	78.26%	78.26%
Inceptionv3	80.17%	76.09%	69.57%
Resnte50	81.4%	70.25%	72.15%
Darknet53	83.47%	79.35%	80.65%

In the context of our experimental investigations, the previously mentioned process was implemented for training and testing for five iterations, while employing a data partitioning strategy that allocated 65% for training and 35% for validation. The main objective was again to detect objects within three distinct categories: planes, ships, and trains. Moreover, this study seeks to gauge the performance of our new novel hybrid object detection algorithm by comparing it with established models, namely the YOLOV7 and DETR. Table 2 displays measurements that encompass critical evaluation metrics, averaged over five iterations of tests for the three aforementioned object categories. Specifically, these metrics are the $mAP_{0.5}$ and $mAP_{0.5:0.95}$ quantities; the accuracy related to the training of the CNN- Resnet101 automatic selection module; the "gain", which refers to the absolute value of the disparity between the correct detection scores of YOLOv7 and DETR; and the "loss", which indicates the absolute value of the difference between the incorrect detection scores of YOLOv7 and DETR.

These comprehensive evaluations (i.e., metrics) offer insights into the performance of our hybrid structure in comparison to the individual YOLO and DETR conventional methods. Therefore, based on Table 2 as well as Figure 4, where the $mAP_{0.5}$ and $mAP_{0.5:0.95}$ values for each object and method are compared, the superiority of our new approach in terms of efficiency and effectiveness in object detection tasks is highlighted.

Table 2. Comprehensive comparison of key metrics, including mAP, training accuracy, gain, and loss, calculated as averages from five separate test iterations.

Object	$mAP_{0.5}$			$mAP_{0.5:0.95}$			Training Accuracy	Gain	Loss
	YOLOv7	DETR	AS	Yolov7	DETR	AS	CNN-Res101		
Plane	0.97	0.96	0.98	0.70	0.68	0.79	85%	0.292	0.13
Ship	0.96	0.97	0.97	0.54	0.64	0.64	83.7%	0.17	0.16
Train	0.979	0.98	0.98	0.68	0.70	0.74	88.57%	0.1779	0.0697

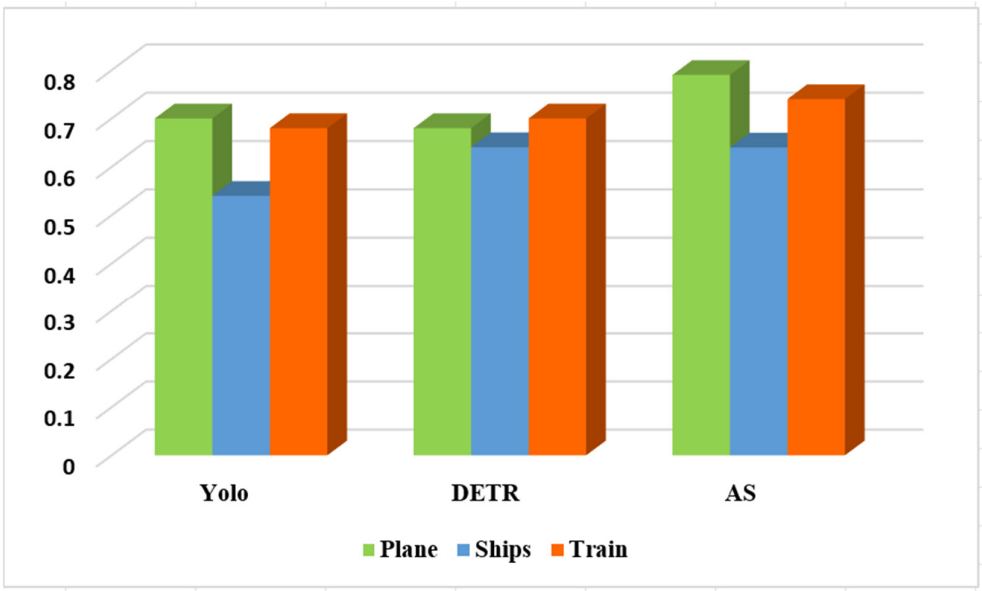

Figure 4. The $mAP_{0.5:0.95}$ scores for three object categories: plane, ship, and train, providing a comparison between the AS module, YOLOv7, and DETR.

4. Discussion

This section outlines the specifics of implementing and training hybrid structures, along with an examination of the results obtained.

4.1. Implementation Details

In this work, we trained a new AS module using a dataset comprising 394 images. This module's input layer operated on standardized images with dimensions set at 224 × 224 pixels, incorporating three color channels representing the Red, Green, and Blue (RGB) color spectrum. In the training phase, the Adam optimization algorithm is operated with specific hyperparameter choices to strike a balance between stable convergence and efficient learning [47]. The decision to employ the Adam optimizer is rooted in its adaptive nature, combining momentum. Moreover, RMSprop (root mean square propagation) methods were also chosen to be utilized, due to the fact that they often result in faster convergence and improved performance in deep learning tasks [48]. As for the learning rate, it was initialized at 0.0001, a relatively low value, to ensure cautious updates to model parameters, especially in the early stages of training. In fact, this mitigates the risk of overfitting to the training data [49]. Additionally, the model was trained using information from 30 epochs with the purpose of allowing it to learn and adapt to the underlying patterns at hand. On another note, for each epoch, the data were partitioned into four mini-batches to balance the computational efficiency with stable convergence [50]. Smaller mini-batch sizes are favored in situations wherein computational resources are constrained, as they allow for more frequent updates of model parameters. In fact, this choice contributes to a more stable and efficient optimization process [51]. Moreover, these hyperparameter

selections were made through a combination of established best practices and empirical experimentation, tailored to the specific requirements and characteristics of our dataset and research objectives. Ultimately, they collectively contribute to a well-balanced and effective training regimen for our neural network model.

In the context of object detection, our research focuses on the training of YOLOv7 and DETR models. We initiated the optimization of model parameters with an initial learning rate of 0.0001 and a batch size set to 2. On top of that, to guard against overfitting, we implemented data augmentation methods, which encompassed actions such as image flipping, rotation, and adjustments to hue.

In Table 3, we provide a comparison of the number of epochs required to train both YOLO and DETR-based models across various datasets, focusing on three object categories: planes, ships, and trains. Intriguingly, the results demonstrate that the DETR achieves convergence with fewer epochs compared to YOLOv7, something that showcases its efficiency in terms of training duration. Also, it is worth highlighting that this efficiency is notable, despite DETR having a higher number of parameters in its transformer architecture compared to YOLOv7.

Table 3. Number of epochs and datasets.

Object	Number of Epochs		Number of Images	
	Yolov7	DETR	Training	Validation
Plane	100	20	1965	281
Ships	100	20	456	114
Train	50	20	198	49

4.2. Analysis and Visualization of Object Detection

Within the realm of computer vision and object detection, the evaluation metrics $mAP_{0.5}$ and $mAP_{0.5:0.95}$ serve as pivotal benchmarks for assessing the performance of detection models [52]. These metrics are commonly applied in conjunction with widely recognized datasets like Pascal VOC and COCO, offering robust means to rigorously evaluate the efficacy of object detection algorithms. Specifically, $mAP_{0.5}$ calculates the average precision for each object class at an IoU threshold of 0.5, and subsequently computes the mean across all classes. This metric essentially measures the model's competence in accurately localizing objects when there exists a minimum 50% overlap between the predicted bounding box and the ground truth bounding box. In simpler terms, it quantifies the model's capacity to detect objects that align reasonably well with the ground truth bounding boxes. Conversely, $mAP_{0.5:0.95}$ offers a more comprehensive evaluation by considering a range of IoU thresholds, spanning from 0.5 to 0.95 in small increments (e.g., 0.5, 0.55, 0.6, . . ., 0.95). This metric computes the average precision for each class at each of these IoU thresholds and subsequently calculates the mean across all classes and IoU thresholds. In practical applications, $mAP_{0.5:0.95}$ is typically favored, since it provides a holistic assessment of a model's performance across diverse IoU thresholds, thereby rewarding models capable of detecting objects with varying degrees of spatial overlap with the ground truth. Both $mAP_{0.5}$ and $mAP_{0.5:0.95}$ play a critical role in evaluating object detection models, offering valuable insights into their ability to strike a balance between recall (detecting objects) and precision (accurately detecting objects) at different levels of object overlap with the ground truth.

The experimental results yielded compelling evidence to validate the efficacy of our new approach in enhancing object detection performance. Notably, the discrepancies observed between $mAP_{0.5}$ and $mAP_{0.5:0.95}$ underscored the inherent challenges faced by detection models when confronted with specific IoU thresholds, spanning from 0.5 to 0.95. Furthermore, these disparities manifested uniquely across the three object classes, namely planes, ships, and trains. In fact, a comprehensive analysis of the dataset depicted in Figure 5 revealed the distribution of the number that each of the three objects was detected with respect to the area. Specifically, it showed that planes were predominantly

concentrated within the small and medium area categories, a scenario conducive to YOLO's strengths, owing to its smaller local receptive field. In contrast, ships exhibited a distribution encompassing small, medium, and large area categories, favoring DETR's performance, which surpassed that of YOLO. Conversely, trains were associated with larger areas, aligning perfectly with DETR's capabilities, and consequently resulting in superior performance.

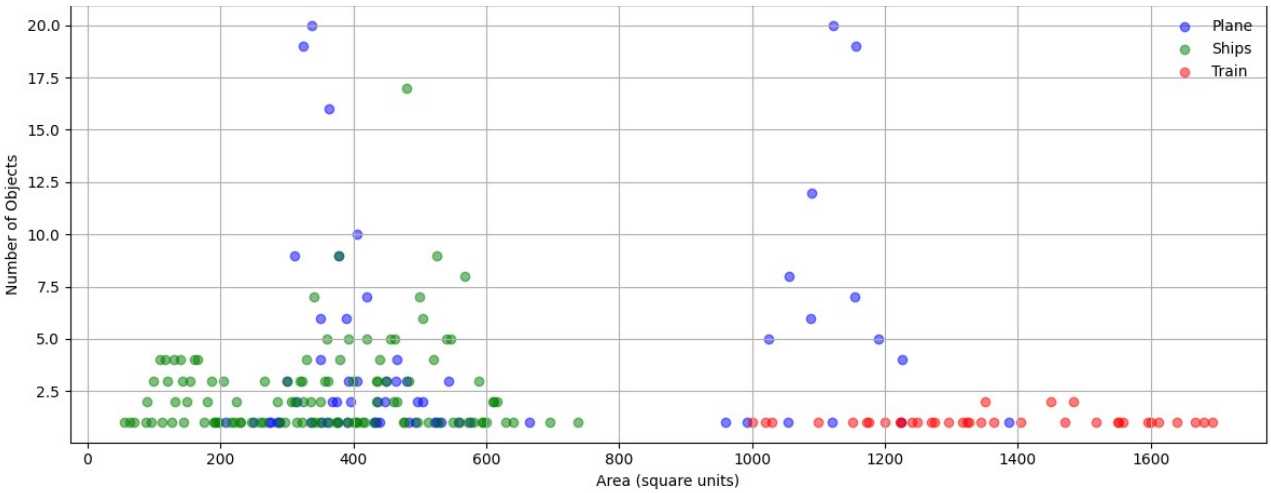

Figure 5. Scatter plot depicting the relationship between area and number of objects.

In conclusion, these findings underscored YOLO's proficiency in handling small objects, DETR's excellence in detecting large objects, and the substantial performance enhancement achieved by the proposed hybrid structures in terms of both the $mAP_{0.5}$ and $mAP_{0.5:0.95}$ metrics. As a result, this improvement in detection and localization capabilities conclusively validates the effectiveness of the proposed approach in the context of object detection compared with individual detection models.

In Figure 6, we visually show the results of our hybrid proposed object detection model for the three distinct object categories: small planes (a), ships detected from multispectral (MS) images (b), ships detected from PANchromatic (PAN) images (c), and train (d). Our primary aim with this visualization is to assess our detection model's performance across a diverse range of objects and image types. Particularly noteworthy is Figure 5a, which emphasizes the hybrid model's efficiency in detecting small planes within intricate settings, including residential areas. In fact, this underscores the model's ability to adapt to varying and challenging contexts. Additionally, Figure 6b,c show ship detection using both multispectral and panchromatic images, providing insights into how the model performs across different imaging modalities. Overall, these visual depictions offer a qualitative perspective on the accuracy and effectiveness of our object detection approach, shedding light on its capacity to identify objects of interest within diverse contextual and imaging scenarios.

Based on the aforementioned, our comprehensive evaluation not only reaffirms the effectiveness of the proposed hybrid object detection model in terms of accuracy and efficiency. On top of that, we explored the model's performance on datasets of varying sizes and complexities, thus addressing potential concerns related to scalability. In fact, scalability is essential for accommodating the increasing demand for object detection in large-scale and dynamic environments. With this in mind, our results revealed that our proposed hybrid approach achieves a consistent and robust performance across diverse datasets. Finally, these findings contribute valuable insights to the broader field of computer vision, emphasizing the model's versatility and applicability in real-world scenarios.

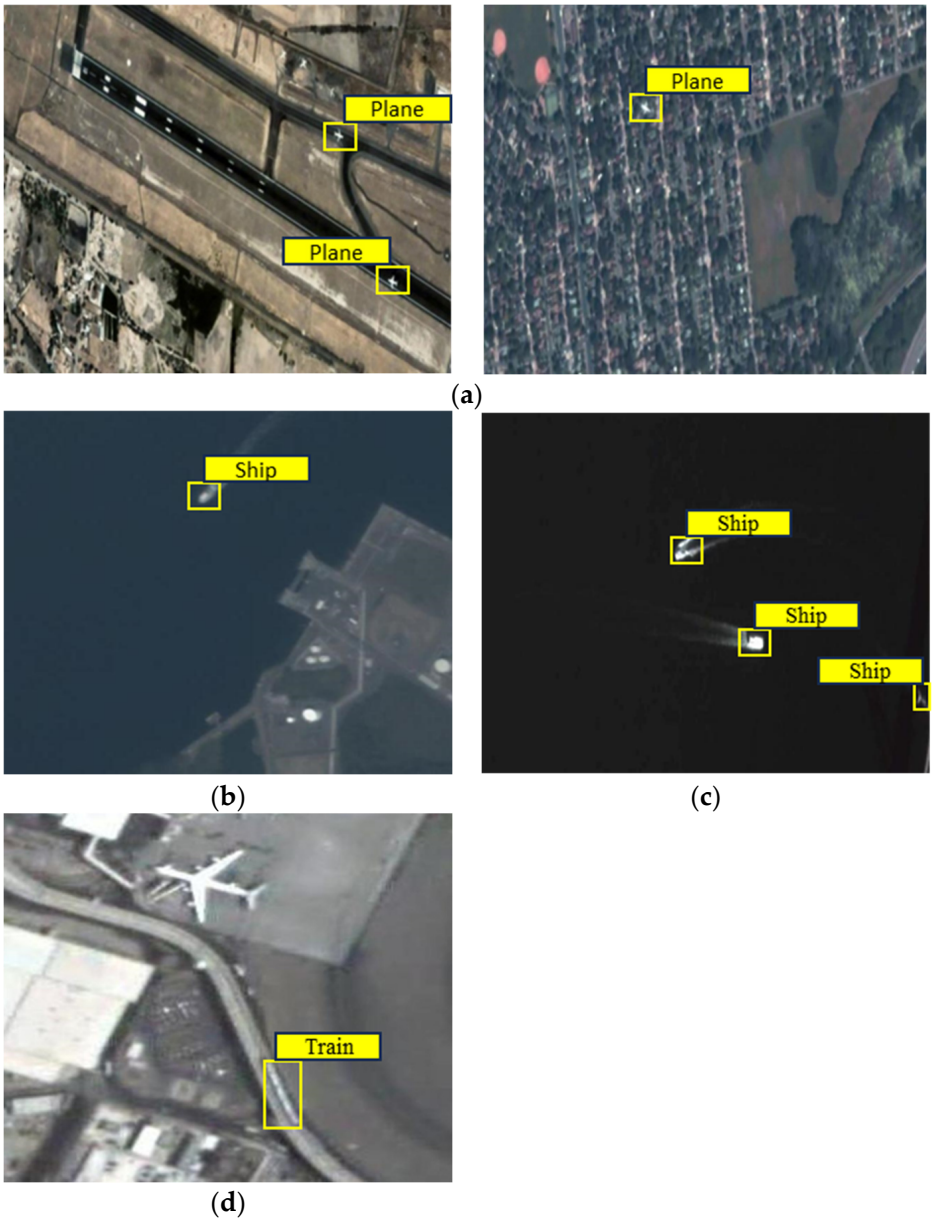

Figure 6. Detection of three objects: (**a**) small planes; (**b**) ships from MS images; (**c**) ships from PAN images; and (**d**) trains.

5. Conclusions

In this paper, we introduced an innovative hybrid approach for object detection in remote sensing applications, which combines the strengths of YOLOv7 and DETR. This integration effectively strikes a balance between local object detection precision and global context awareness, resulting in a substantial improvement in object detection accuracy and enhanced object localization precision.

One of the key advantages of our approach is the inclusion of an automatic selection module. This module, trained using detection accuracy scores based on the mean average precision $mAP_{0.5:0.95}$, serves as an intelligent decision-making component. It optimizes the choice between YOLOv7, and detection transformers based on the unique characteristics of each image, further enhancing object localization accuracy.

Our experimental results undeniably demonstrate the superior performance of our hybrid approach, establishing it as a top-performing solution for remote sensing object detection. This research not only pushes the boundaries of the field, but also holds signifi-

cant promise for applications in geospatial information extraction, land use classification, disaster monitoring, and infrastructure planning within the realm of remote sensing. Ultimately, our approach not only enhances object detection, but also refines the precision of object localization in remote sensing applications, marking a substantial advancement in this domain.

Author Contributions: Conceptualization, M.A.; methodology, A.M.; software, M.A. and A.M.; validation, M.A.; formal analysis, A.M. and M.A.; investigation, M.A.; data curation, M.A.; writing—original draft preparation, M.A.; writing—review and editing, N.E.-S., H.L. and A.M.; visualization, M.A.; supervision, N.E.-S. and H.L.; funding acquisition, N.E.-S. All authors have read and agreed to the published version of the manuscript.

Funding: This project was partially funded by a scholarship to the first author from the Egyptian government, and the Canadian NSERC project number RT691875 of Naser El-Sheimy.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank Chrysostomos Minaretzis for his constructive feedback and valuable help.

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

References

1. Mou, L.; Ghamisi, P.; Zhu, X.X. Deep Recurrent Neural Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3639–3655. [CrossRef]
2. Feng, H.-Z.; Yu, H.-Y.; Wang, W.-Y.; Wang, W.-X.; Du, M.-Q. Recognition of mortar pumpability via computer vision and deep learning. *J. Electron. Sci. Technol.* **2023**, *21*, 100215. [CrossRef]
3. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.-S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [CrossRef]
4. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; Volume 8689, pp. 818–833. [CrossRef]
5. Ziegler, T.; Fritsche, M.; Kuhn, L.; Donhauser, K. Efficient Smoothing of Dilated Convolutions for Image Segmentation. *arXiv* **2019**, arXiv:1903.07992.
6. Lin, M.; Chen, Q.; Yan, S. Network In Network. In Proceedings of the 2nd International Conference on Learning Representations (ICLR)-Conference Track, Banff, AB, Canada, 14–16 April 2014.
7. 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 9th International Conference on Learning Representations (ICLR), Online, 3–7 May 2021.
8. Zhao, Z.; Hu, D.; Wang, H.; Yu, X. Convolutional Transformer Network for Hyperspectral Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6009005. [CrossRef]
9. Yang, X.; Cao, W.; Lu, Y.; Zhou, Y. Hyperspectral Image Transformer Classification Networks. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5528715. [CrossRef]
10. Aleissaee, A.A.; Kumar, A.; Anwer, R.M.; Khan, S.; Cholakkal, H.; Xia, G.-S.; Khan, F.S. Transformers in Remote Sensing: A Survey. *Remote Sens.* **2023**, *15*, 1860. [CrossRef]
11. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the 16th European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M., Eds.; Springer: Cham, Switzerland; pp. 213–229. [CrossRef]
12. 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; pp. 779–788. [CrossRef]
13. Zhenyu, H. Research on Small Target Detection in Optical Remote Sensing Based on YOLOv7. In Proceedings of the 2023 IEEE International Conference on Sensors, Electronics and Computer Engineering (ICSECE), Jinzhou, China, 18–20 August 2023; pp. 804–809. [CrossRef]
14. Alganci, U.; Soydas, M.; Sertel, E. Comparative Research on Deep Learning Approaches for Airplane Detection from Very High-Resolution Satellite Images. *Remote Sens.* **2020**, *12*, 458. [CrossRef]
15. Gidaris, S.; Komodakis, N. LocNet: Improving Localization Accuracy for Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 27–30 June 2016; pp. 789–798. [CrossRef]
16. Li, Q.; Chen, Y.; Zeng, Y. Transformer with Transfer CNN for Remote-Sensing-Image Object Detection. *Remote Sens.* **2022**, *14*, 984. [CrossRef]

17. Shih, K.-H.; Chiu, C.-T.; Lin, J.-A.; Bu, Y.-Y. Real-Time Object Detection With Reduced Region Proposal Network via Multi-Feature Concatenation. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *31*, 2164–2173. [CrossRef]
18. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271. [CrossRef]
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
20. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All you Need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008. [CrossRef]
22. Zhang, X.; Wei, F.; Zhou, M. HIBERT: Document Level Pre-training of Hierarchical Bidirectional Transformers for Document Summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 December 2019; Korhonen, A., Traum, D., Màrquez, L., Eds.; Association for Computational Linguistics: Toronto, ON, Canada; pp. 5059–5069. [CrossRef]
23. Zhang, Z.; Lu, X.; Cao, G.; Yang, Y.; Jiao, L.; Liu, F. ViT-YOLO: Transformer-Based YOLO for Object Detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 11–17 October 2021; pp. 2799–2808. [CrossRef]
24. Cesar, L.B.; Manso-Callejo, M.-Á.; Cira, C.-I. BERT (Bidirectional Encoder Representations from Transformers) for Missing Data Imputation in Solar Irradiance Time Series. *Eng. Proc.* **2023**, *39*, 26. [CrossRef]
25. Yu, W.; Yang, T.; Chen, C. Towards Resolving the Challenge of Long-Tail Distribution in UAV Images for Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 3–8 January 2021; pp. 3258–3267. [CrossRef]
26. Zoph, B.; Ghiasi, G.; Lin, T.-Y.; Cui, Y.; Liu, H.; Cubuk, E.D.; Le, Q. Rethinking Pre-training and Self-training. In Proceedings of the Advances in Neural Information Processing Systems, Online, 6–12 December 2020; Volume 33, pp. 3833–3845. [CrossRef]
27. Wu, S.; Li, X.; Wang, X. IoU-aware single-stage object detector for accurate localization. *Image Vis. Comput.* **2020**, *97*, 103911. [CrossRef]
28. Wang, T.; Lin, Q. Hybrid predictive models: When an interpretable model collaborates with a black-box model. *J. Mach. Learn. Res.* **2021**, *22*, 6085–6122. [CrossRef]
29. Manogaran, G.; Lopez, D. A survey of big data architectures and machine learning algorithms in healthcare. *Int. J. Biomed. Eng. Technol.* **2017**, *25*, 182–211. [CrossRef]
30. Zhang, Q.; Wu, Y.N.; Zhu, S.-C. Interpretable Convolutional Neural Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (ICCVW), Lake City, UT, USA, 18–23 June 2018; pp. 8827–8836. [CrossRef]
31. Xu, X.; Feng, Z.; Cao, C.; Li, M.; Wu, J.; Wu, Z.; Shang, Y.; Ye, S. An Improved Swin Transformer-Based Model for Remote Sensing Object Detection and Instance Segmentation. *Remote Sens.* **2021**, *13*, 4779. [CrossRef]
32. Liu, X.; Ma, S.; He, L.; Wang, C.; Chen, Z. Hybrid Network Model: TransConvNet for Oriented Object Detection in Remote Sensing Images. *Remote Sens.* **2022**, *14*, 2090. [CrossRef]
33. Li, Q.; Zhong, R.; Du, X.; Du, Y. TransUNetCD: A Hybrid Transformer Network for Change Detection in Optical Remote-Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5622519. [CrossRef]
34. Wang, H.; Chen, X.; Zhang, T.; Xu, Z.; Li, J. CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. *Remote Sens.* **2022**, *14*, 1956. [CrossRef]
35. Shen, X.; Liu, B.; Zhou, Y.; Zhao, J. Remote sensing image caption generation via transformer and reinforcement learning. *Multimed. Tools Appl.* **2020**, *79*, 26661–26682. [CrossRef]
36. 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 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475. [CrossRef]
37. Makanapura, N.; Sujatha, C.; Patil, P.R.; Desai, P. Classification of plant seedlings using deep convolutional neural network architectures. *J. Phys. Conf. Ser.* **2022**, *2161*, 012006. [CrossRef]
38. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826. [CrossRef]
39. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
40. Xu, Z.; Sun, K.; Mao, J. Research on ResNet101 Network Chemical Reagent Label Image Classification Based on Transfer Learning. In Proceedings of the 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Weihai, China, 9 March 2021; pp. 354–358. [CrossRef]
41. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
42. Huang, J.; Rathod, V.; Sun, C.; Zhu, M.; Korattikara, A.; Fathi, A.; Fischer, I.; Wojna, Z.; Song, Y.; Guadarrama, S.; et al. Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 9 November 2017; pp. 3296–3297. [CrossRef]

43. Li, Y.; Mao, H.; Girshick, R.; He, K. Exploring Plain Vision Transformer Backbones for Object Detection. In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T., Eds.; Springer: Cham, Switzerland; pp. 280–296. [CrossRef]
44. Yang, X.; He, S.; Wu, J.; Yang, Y.; Hou, Z.; Ma, S. Exploring Spatial-Based Position Encoding for Image Captioning. *Mathematics* **2023**, *11*, 4550. [CrossRef]
45. Liu, K.; Sun, Q.; Sun, D.; Peng, L.; Yang, M.; Wang, N. Underwater Target Detection Based on Improved YOLOv7. *J. Mar. Sci. Eng.* **2023**, *11*, 677. [CrossRef]
46. Yin, Q.; Hu, Q.; Liu, H.; Zhang, F.; Wang, Y.; Lin, Z.; An, W.; Guo, Y. Detecting and Tracking Small and Dense Moving Objects in Satellite Videos: A Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5612518. [CrossRef]
47. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR)-Conference Track, San Diego, CA, USA, 7–9 May 2015. [CrossRef]
48. Zeiler, M.D. ADADELTA: An Adaptive Learning Rate Method. *arXiv* **2012**, arXiv:1212.5701.
49. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2017**, arXiv:1609.04747.
50. Gao, Y.; Li, J.; Zhou, Y.; Xiao, F.; Liu, H. Optimization Methods For Large-Scale Machine Learning. In Proceedings of the 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 19 January 2022; pp. 304–308. [CrossRef]
51. LeCun, Y.A.; Bottou, L.; Orr, G.B.; Müller, K.-R. Efficient BackProp. In *Neural Networks: Tricks of the Trade*; Montavon, G., Orr, G.B., Müller, K.-R., Eds.; Lecture Notes in Computer Science (Volume 7700); Springer: Cham, Switzerland; Berlin/Heidelberg, Germany, 2012; pp. 9–48. [CrossRef]
52. Wood, L.; Chollet, F. Efficient Graph-Friendly COCO Metric Computation for Train-Time Model Evaluation. *arXiv* **2022**, arXiv:2207.12120.