remote sensing

MDPI

Article

Transformer with Transfer CNN for Remote-Sensing-Image Object Detection

Qingyun Li, Yushi Chen * and Ying Zeng

School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150001, China; 21b905003@stu.hit.edu.cn (Q.L.); 1180500504@stu.hit.edu.cn (Y.Z.)
* Correspondence: chenyushi@hit.edu.cn

Abstract: Object detection in remote-sensing images (RSIs) is always a vibrant research topic in the remote-sensing community. Recently, deep-convolutional-neural-network (CNN)-based methods, including region-CNN-based and You-Only-Look-Once-based methods, have become the de-facto standard for RSI object detection. CNNs are good at local feature extraction but they have limitations in capturing global features. However, the attention-based transformer can obtain the relationships of RSI at a long distance. Therefore, the Transformer for Remote-Sensing Object detection (TRD) is investigated in this study. Specifically, the proposed TRD is a combination of a CNN and a multiple-layer Transformer with encoders and decoders. To detect objects from RSIs, a modified Transformer is designed to aggregate features of global spatial positions on multiple scales and model the interactions between pairwise instances. Then, due to the fact that the source data set (e.g., ImageNet) and the target data set (i.e., RSI data set) are quite different, to reduce the difference between the data sets, the TRD with the transferring CNN (T-TRD) based on the attention mechanism is proposed to adjust the pre-trained model for better RSI object detection. Because the training of the Transformer always needs abundant, well-annotated training samples, and the number of training samples for RSI object detection is usually limited, in order to avoid overfitting, data augmentation is combined with a Transformer to improve the detection performance of RSI. The proposed T-TRD with data augmentation (T-TRD-DA) is tested on the two widely-used data sets (i.e., NWPU VHR-10 and DIOR) and the experimental results reveal that the proposed models provide competitive results (i.e., centuple mean average precision of 87.9 and 66.8 with at most 5.9 and 2.4 higher than the comparison methods on the NWPU VHR-10 and the DIOR data sets, respectively) compared to the competitive benchmark methods, which shows that the Transformer-based method opens a new window for RSI object detection.

Keywords: convolutional neural network (CNN); object detection; remote-sensing images; transfer learning; Transformer

Citation: Li, Q.; Chen, Y.; Zeng, Y. Transformer with Transfer CNN for Remote-Sensing-Image Object Detection. *Remote Sens.* **2022**, *14*, 984. https://doi.org/10.3390/rs14040984

Academic Editors: Xin Wu, Danfeng Hong, Sicong Liu and Pedram Ghamisi

Received: 22 January 2022
Accepted: 13 February 2022
Published: 17 February 2022

1. Introduction

Object detection in remote-sensing images (RSIs) is used to answer one of the most basic questions in the remote-sensing (RS) community: What and where are the objects (such as a ship, vehicle, or aircraft) in the RSIs? In general, the objective of object detection is to build models to localize and recognize different ground objects of interest in high-resolution RSIs [1]. Due to the fact that object detection is a fundamental task for the interpretation of high-resolution RSIs, a great number of methods have been proposed to handle the issue of RSI object detection in the last decade [2].

The traditional RSI object-detection methods focus on constructing effective features for objects of interest and training a classifier from a set of annotated RSIs. They usually acquire object regions with sliding windows and then try to recognize each region. The varieties of feature-extracting methods, e.g., bag-of-words (BOW) [3], scale-invariant feature transform [4], and their extensions, have been explored for representing objects. Then, the

feature fusion and dimension processing were conducted in order to further improve the representation capability of multiple features. At last, efficient and well-designed classifiers were trained to recognize objects. For example, Sun et al. [5] proposed an RSI detection framework based on spatial sparse coding bag-of-words (SSCBOW), which adopted a rotation-invariant spatial-mapping strategy and sparse coding to decrease reconstruction error. Cheng et al. [6] explored a partially model-based RSI object-detection method based on the collection of part detectors (COPD), which used linear support-vector machines (SVMs) as partial models for detecting objects or recurring patterns. These methods could be adapted to more complicated tasks, but the hand-crafted feature-extracting approaches significantly restricted the detection performance.

As the detection performance of methods based on hand-crafted features and inefficient region-proposal strategies became saturated, it was hard to make substantial progress on object-detection until the emergence of deep convolutional neural networks (CNNs) [7]. Relying on the ability of CNNs to extract high-level and robust features, Girshick et al. [8,9] proposed region-CNN (R-CNN) and Fast R-CNN, which achieved an attractive detection performance. These methods used CNN to classify and locate objects from a specified amount of generated region proposals (bounding-box candidates). Subsequently, numerous researchers explored RSI object-detection methods based on the R-CNN framework. Cheng et al. [10] inserted a fully connected layer into the tail of the backbone network of the R-CNN framework and restrained the inserted layer with a regularization constraint to minimize the rotation variation. Thus, a rotation-invariant CNN (RICNN) was constructed. Afterward, a fisher discrimination regularized layer was appended to construct an enhanced RICNN, i.e., the RIFD-CNN [11]. Inspired by the idea of the region-proposal network (RPN) in the Faster R-CNN [12], Li et al. [13] presented multi-angle anchors for establishing a rotation-insensitive RPN, and a double-channel network was used for contextual feature fusion. The utilization of the RPN enormously reduced the time for region proposal and achieved a near-real-time speed. Additionally, to enhance the detection performance for small objects in RSIs, some researchers began to develop RSI object-detection methods based on multi-scale feature operations. Inspired by the feature-pyramid network (FPN) [14], Zhang et al. [15] presented a double multi-scale FPN framework and studied several multi-scale training and inference strategies. Deng et al. [16] and Guo et al. [17] focused on multi-scale object-proposal networks that generated candidate regions with features of different intermediary layers, and the multi-scale object-detection networks made predictions on the obtained regions. The R-CNN-based RSI object-detection methods made great progress on detection performance, but they still suffered from insufficient inference speeds caused by redundant computations.

Methods based on the framework of the R-CNN always obtained region proposals first and then predicted categories and refined their coordinates; therefore, they were called two-stage RSI object-detection algorithms. In contrast, many researchers focused on exploring methods that complete the whole detection in only one step, which were called one-stage RSI object-detection algorithms [2]. Plenty of these methods were based on one of the most representative studies in the field of object detection, the You Only Look Once (YOLO) [18], which is an extremely fast object-detection paradigm. The YOLO discarded the process of seeking region proposals and directly predicted both bounding-box coordinates and categories, which dramatically accelerated the inference process [18–20]. Pham et al. [21] proposed YOLO-fine, which conducted finer regression in order to enhance the capacity of recognizing small objects and tackled the problems of domain adaptation by investigating its robustness to various backgrounds. Alganci et al. [22] provided a comparison among YOLO-v3 and other CNN-based detectors for RSI and evaluated that the YOLO provided the most balanced trade-off between detection accuracy and computation efficiency. Additionally, a few studies shared similar ideas with the single-shot multibox detector [23]. Zhuang et al. [24] applied a single-shot framework, which focused on multi-scale feature fusion and improved performance for detecting small objects. In general, the one-stage-RSI detection methods were more appropriate for real-time object-detection

tasks. However, it seems that CNN-based methods, whether one-stage or two-stage, have reached the bottleneck of progress.

Recently, the attention-based Transformer presented by Vaswani et al. [25] has become the standard model for machine translation. Numerous studies have demonstrated that the Transformer might also be efficient at image-processing tasks, and they have achieved breakthroughs. The Transformer was able to obtain the relationship in RSIs at a long distance [26–28], which tackled the difficulty of CNN-based methods for capturing global features. Therefore, there have been a number of successful studies focusing on Transformer-based models in the RS community. Inspired by the Vision Transformer [26], He et al. [29] proposed a Transformer-based hyperspectral image-classification method. They introduced the spatial-spectral Transformer, using a CNN to extract spatial features of hyperspectral images and a densely connected Transformer to learn the spectra relationships. Hong et al. [30] presented a flexible backbone network for hyperspectral images named SpectralFormer, which exploited the spectral-wise sequence attributes of hyperspectral images in order to sequentially feed them into the Transformer. Zhang et al. [31] proposed a Transformer-based method for a remote-sensing scene-classification method, which designed a new bottleneck based on multi-head self-attention (MHSA) for image embedding, and cascaded encoder blocks to enhance accuracy. They all achieved state-of-the-art performance, which shows the potential of the Transformer for various tasks in RSI processing. However, for RSI object detection, the amount of studies working on the basis of the Transformer is still insufficient. Zheng et al. [32] proposed an adaptive, dynamically refined one-stage detector based on the feature-pyramid Transformer, which embedded a Transformer in the FPN in order to enhance its feature-fusion capacity. Xu et al. [33] proposed a local-perception backbone based on the Swin Transformer for RSI object detection and instance segmentation, and they investigated the performance of their backbone on different detection frameworks. In their studies, the Transformer worked as a feature-interaction module, i.e., backbone or feature-fusion component, which is adaptable to various detection frameworks. Above all, since the Transformer has enormous potential to promote a unification of the architecture of various tasks in artificial intelligence, it is essential to further explore Transformer-based RSI object detectors.

In this paper, we investigate a neoteric Transformer-based remote-sensing object-detection (TRD) framework. The proposed TRD is inspired by the detection Transformer [28], which takes features obtained from a CNN backbone as the input and directly outputs a set of detected objects. The existing Transformer-based RSIs object detectors [32,33] are still highly dependent on the existing detection frameworks composed of various surrogate-task components, such as duplicated prediction elimination, etc. The proposed TRD abandons the conventional complicated structure in favor of an independent and more end-to-end framework. Additionally, the CNN backbone in the TRD is trained with transfer learning. To reduce the diversity of the source domain and target domain, the T-TRD is proposed, which adjusts the pre-trained CNN with the attention mechanism for a better transfer. Moreover, since the quantity of reliable training samples for RSI object detection is usually insufficient for training a Transformer-based model, the T-TRD-DA explores data augmentation composed of sample expansion and multiple-sample fusion to enrich the training samples and prevent overfitting. We hope that our research will inspire the development of RSI object-detection components based on the Transformer.

In summary, the following are the main contributions of this study.

(1) An end-to-end Transformer-based RSI object-detection framework, TRD, is proposed, in which the Transformer is remolded in order to efficiently integrate features of global spatial positions and capture relationships of feature embeddings and objects instances. Additionally, the deformable attention module is introduced as an essential component of the proposed TRD, which only attends to a sparse set of sampling features and mitigates the problem of high computational complexity. Hence, the TRD can process RSIs on multiple scales and recognize objects of interest from RSIs.

(2) The pre-trained CNN is used as the backbone for feature extraction. Furthermore, in order to mitigate the difference between the two data sets (i.e., ImageNet and RSI data set), the attention mechanism is used in the T-TRD to reweight the features, which further improves the RSI detection performance. Therefore, the pre-trained backbone is better transferred and obtains discriminant pyramidal features.

(3) Data augmentations, including sample expansion and multiple-sample fusion, are used to enrich the diversity of orientations, scales, and backgrounds of training samples. In the proposed T-TRD-DA, the impact of using insufficient training samples for Transformer-based RSI object detection is alleviated.

2. The Proposed Transformer-Based RSI Object-Detection Framework

Figure 1 shows the overview architecture of the proposed Transformer-based RSI object-detection framework. First, a CNN backbone with attention-based transferring learning is used for extracting multi-scale feature maps of the RSIs. The feature maps from the shallower layers have higher resolutions, which benefits the detection of small-object instances, while the high-level features have wide receptive fields and they are appropriate for large-object detection and global spatial-information fusion. The features of all levels are embedded together in a sequence. The sequence of embedded features undergoes the encoder and decoder of the Transformer-based detection head and is transferred to a set of predictions with categories and locations. As the figure shows, the point in the input embeddings from the high-level feature map tends to recognize a small instance, while that from the low-level map is inclined to recognize a large instance. The detailed introduction of the proposed Transformer-based RSI object-detection framework is started with the framework of the proposed TRD and the effective deformable attention module in its Transformer. Subsequently, the attention-based transferring backbone and the data augmentation are introduced in detail.

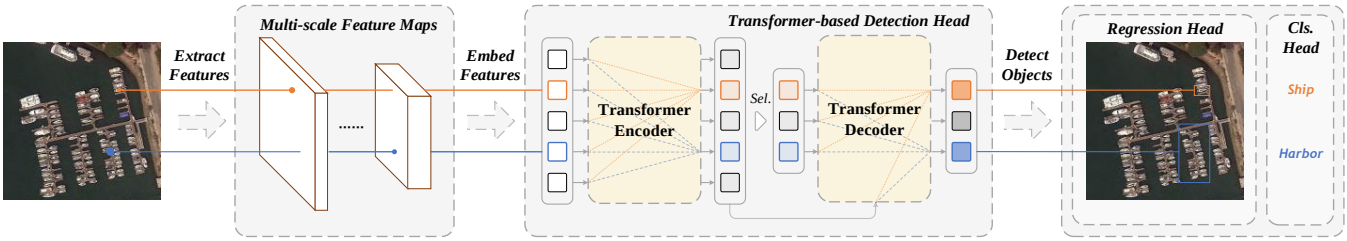


Figure 1. The overview architecture of the proposed Transformer-based RSI object-detection framework.

2.1. The Framework of the Proposed TRD

Figure 2 shows the framework of the proposed TRD. A CNN backbone is first used to extract pyramidal multi-scale feature maps from an RSI. They are then embedded with the 2D positional encoding and converted to a sequence that can be inputted into the Transformer. The Transformer is remolded in order to process the sequence of image embeddings and make predictions of detected object instances.

The feature pyramid of the proposed TRD can be obtained by a well-designed CNN, and in this study, the detection backbone based on ResNet [34] is adopted. The convolutional backbone takes an RSI $I \in \mathbb{R}^{3\times H_0\times W_0}$ of an arbitrary size $H_0 \times W_0$ as the input and generates hierarchical feature maps. Specifically, the ResNet generate hierarchical maps from the outputs of the last three stages, which are denoted as $\{f_1, f_2, f_3\}$, and $f_l \in \mathbb{R}^{C_l\times H_l\times W_l}$. Those of the other stages are not included due to their restricted receptive field and additional computational complexities. Then, the feature map at each level undergoes 1×1 convolutions, mapping their channels C_l to a smaller, uniform dimension d.

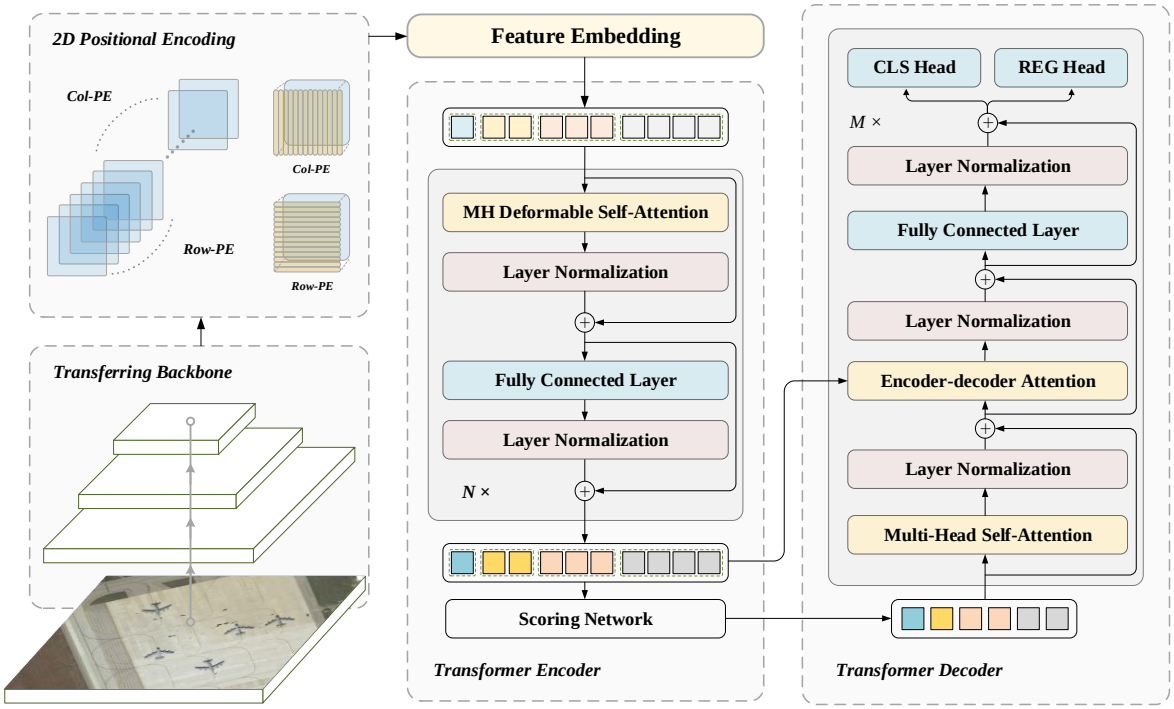


Figure 2. The framework of the proposed TRD.

Hence, a three-level feature pyramid is obtained, which is denoted as $\{x_1, x_2, x_3\}$ and $x_l \in \mathbb{R}^{d \times H_l \times W_l}$. Additionally, a lower-resolution feature map x_4 is acquired by a 3×3 convolution on x_3.

The feature pyramid is further processed to be fed into the Transformer. The MHSA in Transformer aggregates the elements of the input and does not discriminate their positions; hence, the Transformer has permutation invariance. To alleviate this problem, we need to embed spatial information in the feature maps. Therefore, after the L-level feature pyramid $\{x_l\}_{l=1}^{L}$ is extracted from the convolutional backbone, the 2D position encodings are supplemented at each level. Specifically, the sine and cosine positional encoding of the original Transformer is extended to column and row positional encodings, respectively. They are both acquired by encoding on the dimension of the row or column and half of the d channels, and then duplicated to the other spatial dimension. The final positional encodings are concatenated with them.

The Transformer expects a sequence consisting of elements of equal dimensions as inputs. Therefore, the multi-scale position-encoded feature maps $\{x_l\}_{l=1}^{L}$ are flattened in the spatial dimensions, developing them into L sequences of $H_l \times W_l$ length. The input sequence is obtained by concatenating the sequences from L levels, which consists of $\sum_{l=1}^{L} H_l \times W_l$ tokens with d dimensionalities. Each pixel in the feature pyramid is treated as an element of the sequence. The Transformer then models the interaction of the feature points and recognizes concerned object instances from the sequence.

The original Transformer adopted an encoder–decoder structure using stacked self-attention layers and point-wise fully connected layers, and the decoder was auto-regressive, generating an element at a time and appending the element to the input sequence for the next generation [25]. In a different manner, the Transformer here changes the MHSA layers of the encoder to the deformable attention layers, which are more attractive for modeling the relationship between feature points due to the lack of computational and memory complexities. Besides, the decoder adopts a non-autoregressive structure, which parallelly decodes the elements. The details are as follows:

The encoder takes the sequence of the feature embeddings as the input and outputs a sequence of spatial-aware elements. The encoder consists of N cascaded encoder layers. In each encoder layer, the sequence undergoes a deformable multi-head attention layer

and a feed-forward layer, both of which are accompanied by a layer normalization and a residual computation, and the encoder layer outputs an equilong sequence of isometric elements. The deformable attention layers aggregate the features at positions in an adaptive field, obtaining feature maps with a distant relationship. The feature points can be used to compose the input sequence of the decoder. To reduce computational complexities, the feature points are fed into a scoring network, specifically, a three-layer FFN with a softmax layer, which can be realized as a binary classifier of the foreground and background. The N_p highest scored points constitute a fixed-length sequence, which is fed into the decoder. The encoder endows the multi-scale feature maps with global spatial information and then selects a quantity-fixed set of spatial-aware feature points, which are more easily used for detecting object instances.

The decoder takes the sequence of essential feature points as the input and outputs a sequence of object-aware elements in parallel. The decoder also contains M cascaded decoder layers, consisting of an MHSA layer, an encoder–decoder attention layer, and a feed-forward layer, followed by three-layer normalization and residual computations behind them, respectively. The MHSA layers capture interactions between pairwise feature points, which has benefits for constraints related to object instances, such as preventing duplicate predictions. Each encoder–decoder attention layer takes the elements from the previous layer in the decoder as queries and those from the output of the last encoder layer as memory keys and values. It enables the feature points to attend to feature contexts at different scale levels and global spatial positions. The output embeddings of each decoder layer are fed into a layer normalization and the prediction heads, which share a common set of parameters for different layers.

The prediction heads further decode the output embeddings from the decoder into object categories and bounding-box coordinates. Similar to most modern end-to-end object-detection architectures, the prediction head is divided into two branches for classification and regression. In the classification branch, a linear projection with a softmax function is used to predict the category of each embedding. A special 'background' category is appended to the classes, meaning that no concerned object is detected from the query. In the regression branch, a three-layer fully connected network with the ReLU function is utilized for producing the normalized coordinates of the bounding boxes. In total, the heads generate an N_p set of predictions, and each set consists of a class and the corresponding box position. The final prediction results are obtained by removing the 'background'.

The proposed TRD takes full advantage of the relationship-capturing capacity of the Transformer and rebuilds the original structure and embedding scheme. It explores a Transformer-based paradigm for RSI object detection.

2.2. The Deformable Attention Module

To enhance the detection performance of small-object instances, the idea of utilizing multi-scale feature maps is explored, in which the low-level and high-resolution feature maps are conducive to recognizing small objects. However, the high-resolution feature maps result in high computational and memory complexities for the conventional MHSA-based Transformer, because the MHSA layers measure the compatibility of each pair of reference points. In contrast, the deformable attention module only pays attention to a fixed-quantity set of essential sampling points at several adaptive positions around the reference point, which enormously decreases the computational and memory complexities. Thus, the Transformer can be effectively extended to the aggregation of multi-scale features of RSIs.

Figure 3 shows the diagram of the deformable attention module. The module generates a specific quantity of sampling offsets and attention weights for each element in each scale level. The features at the sampling positions of maps in different levels are aggregated to a spatial- and scale-aware element.

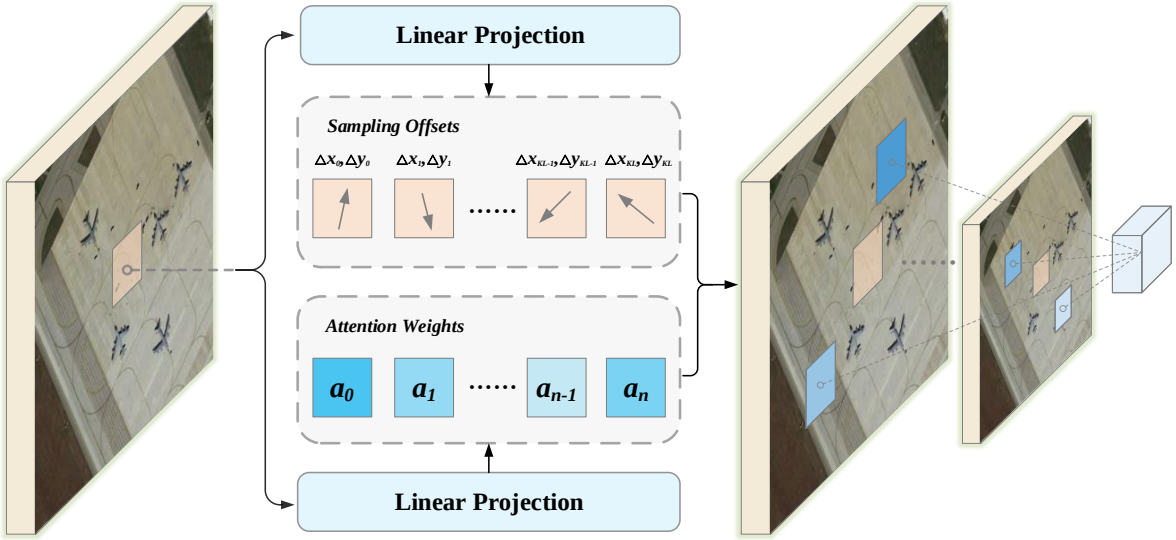


Figure 3. The diagram of the deformable attention module.

The input sequence of the embedded feature elements is denoted as $\boldsymbol{x}$. In each level, the normalized location of the q-th feature element is denoted as $\hat{\boldsymbol{p}}_q \in [0,1]^2$, which can be re-scaled to the practical coordinates at the l-th level with a mapping function $\phi_l\left(\hat{\boldsymbol{p}}_q\right)$. For each element, which is represented as $\boldsymbol{x}\left(\phi_l\left(\hat{\boldsymbol{p}}_q\right)\right)$, a $3LK$-channel linear projection is used to obtain LK sets of sampling offsets $\Delta\boldsymbol{p}_{lkq} \in \mathbb{R}^2$ and attention weights $a_{lkq} \in [0,1]$, which is normalized by $\sum_{l=1}^{L}\sum_{k=1}^{K} a_{lkq} = 1$. Then, the features of the LK sampling points $\boldsymbol{x}\left(\phi_l\left(\hat{\boldsymbol{p}}_q\right) + \Delta\boldsymbol{p}_{lkq}\right)$ are calculated from the input feature maps by applying bilinear interpolation. They are aggregated by multiplying the attention weights a_{lkq}, generating a spatial- and scale-aware element. Therefore, the output sequence of the deformable attention module is calculated with (1).

$$\mathcal{F}(\boldsymbol{x}) = \sum_{l=1}^{L}\sum_{k=1}^{K} \boldsymbol{A}_{lk}\cdot\boldsymbol{W}_v\boldsymbol{x}(\boldsymbol{p}_l + \Delta\boldsymbol{p}_{lk}) \tag{1}$$

where l indexes the L feature levels, and k indexes the K sampled points for keys and values, respectively. The $\boldsymbol{p}_l$ is the sequence of the practical coordinates $\{\phi_l(\hat{\boldsymbol{p}}_0), \phi_l(\hat{\boldsymbol{p}}_1), \cdots\}$, and the $\Delta\boldsymbol{p}_{lk}$ indicates the sequence of the k-th sampling offsets $\{\Delta\boldsymbol{p}_{lk0}, \Delta\boldsymbol{p}_{lk0}, \cdots\}$. The $\boldsymbol{A}_{lk}$ is composed of normalized attention weights a_{lkq}.

The deformable attention mechanism resolves the problem of processing spatial features with self-attention computations. It is extremely appropriate for Transformers in computer-vision tasks and it is adopted in the proposed TRD detector.

2.3. The Attention-Based Transferring Backbone

In general, deep CNN can obtain discriminative features of RSIs for object detection. However, due to the fact that RSI object-detection tasks usually have limited training samples and deep models always contain numerous parameters, deep-learning-based RSI object-detection methods usually face the problem of overfitting.

To address the overfitting issue, transfer learning is used in this study. In the proposed T-TRD detector, a pre-trained CNN model is used as the backbone for RSI feature extraction, and then the Transformer-based detection head is used to complete the object-detection task. In CNNs, the first few convolution operations extract low-level and mid-level features such as blobs, corners, and edges, which are common features for image processing [35].

In RSI object detection, the proper re-usage of low-level and mid-level representations will significantly improve the detection performance. However, due to the fact that the spatial resolution and imaging environment between ImageNet and RSI are quite different,

the attention mechanism is used in this study to adjust the pre-trained model for better RSI object detection.

In the original attention mechanism, more attention is paid to the important regions in an image and the selected regions are assigned by different weights. Such an attention mechanism has been proved to be effective in text entailment and sentence representations [36,37].

Motivated by the attention mechanism, we re-weight the feature maps to reduce the difference in the two data sets (i.e., RSI and ImageNet). Specifically, the feature maps in the pre-trained model are re-weighted and then transferred to the backbone in RSI object detection. When attention scores of different feature maps are higher, the transferring features are more important for the following feature extractions. Figure 4 shows the framework of the proposed attention-based transferring backbone. As is shown, the model pre-trained on the source-domain-images data set is transferred to the backbone of the T-TRD. The attention weights are obtained with global average pooling and non-linear projection. At last, the feature maps are re-weighted according to the attention weights. The detailed steps are defined below.

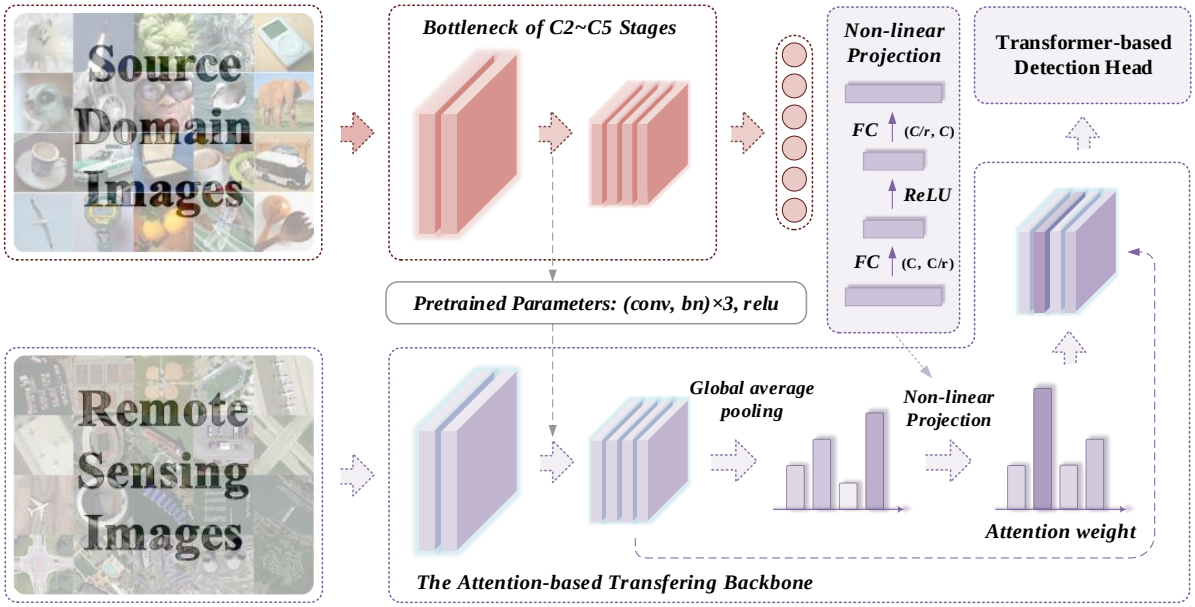


Figure 4. The framework of the proposed attention-based transferring backbone.

At first, feature maps in one convolutional layer are operated to channel-wise statistics by using the global average pooling layer. Specifically, the spatial dimension $H' \times W'$ of each feature map is calculated by the following formula:

$$v = \frac{1}{H' \times W'} \sum_{q=1}^{H'} \sum_{s=1}^{W'} u(q, s) \tag{2}$$

where u refers to the input feature map and v indicates the aggregated information of a whole feature map.

Next, to capture the relationships of feature maps with different importances, a neural network that consists of two fully connected (FC) layers and a ReLU operation are utilized. To limit model complexity, the first FC layer maps the total number of feature maps to a fixed value (i.e., 128), followed by a non-linearity ReLU operation. In addition, the second FC layer restores the number of feature maps to its initial dimension. By learning the parameters in this neural network through backpropagation, the interaction reflected the importance between different feature maps can be obtained.

Finally, the attention values of different feature maps are outputted by the sigmoid function, which restricts the values from zero to one. Each feature map multiplies the obtained attention values to distinguish the degree importance of different feature maps.

The above steps are used in the proposed attention-based transferring backbone. The transferring features from ImageNet to RSI re-weighted by the attention values could boost the feature discriminability, thereby reducing the difference between the two data sets by learning more important transferring features and weakening less important features.

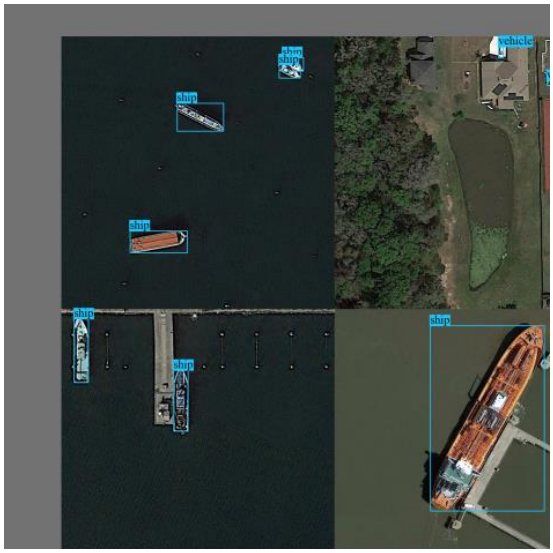
2.4. Data Augmentation for RSI Object Detection

As is reported, the Transformer-based vision models are more likely to overfit than the CNN with equivalent computational complexity on limited data sets [26]. However, the quantities of training samples in RSI data sets for object detection are usually deficient. Additionally, objects in an RSI sample are usually sparsely distributed, which is an inefficient method of training the proposed Transformer-based detection models. Hence, a data-augmentation method, which is composed of sample expansion and multiple-sample fusion, is merged into the training strategy of the T-TRD to improve the detection performance.

Let $X = \{x_1, x_2, \cdots, x_N\}$ be the training samples. We define a set of four right-angle rotation transformations $T_R = \{t_{R0}, t_{R1}, t_{R2}, t_{R3}\}$ and another set of two horizontal flip transformations $T_F = \{t_{F0}, t_{F1}\}$. Both sets are applied to all the training samples, generating a $\times 8$ extended samples set $T_F T_R X = \{t_{F0} t_{R0} x_1, t_{F1} t_{R0} x_1, t_{F0} t_{R1} x_1, \cdots, t_{F1} t_{R3} x_N\}$.

For each sample in the extended set, we randomly choose three samples from the set and blend the four samples into a larger fixed-size sample. The samples are concatenated at the top-left, top-right, bottom-left, and bottom-right of an intersection point. Afterward, a blank canvas of the fusion image size is generated by gray padding. Then, the normalized coordinates of the intersection point are randomly generated, with a restricted range of 0.25 to 0.75. The concatenated sample is pasted on the canvas by aligning the intersection points. The composite images and boxes outside of the border of the canvas are cropped. Figure 5 shows several examples of composite RSI samples. At last, random scale and crop augmentation are applied to the composite samples.

(**a**) A composite example.



(**b**) A composite example.

Figure 5. *Cont.*

(**c**) A composite example.

(**d**) A composite example.

Figure 5. Several examples of composite samples for RSI object detection. The samples are from the DIOR data set, whose size is 800 × 800, and their spatial resolutions are between 30 m and 0.5 m. The size of composite images is set at 1600 × 1600.

With the data augmentation, the problem of insufficient training samples is mitigated. The proposed T-TRD-DA trains a Transformer-based detection model on an enhanced training data set with more diversity of scale, orientation, background, etc., which prevents the proposed deep model from overfitting.

3. Data Sets and Experimental Settings

3.1. Data Description

The proposed TRD, T-TRD and T-TRD-DA are evaluated on the NWPU VHR-10 [6] and DIOR [2] data sets, which are both widely-used public data sets for multi-class object detection in RSIs.

The NWPU VHR-10 data set contains 800 very-high-resolution RSIs collected from Google Earth and the Vaihingen data set [38]. There is an annotated 'positive image set' and a 'negative image set'. The 150 images in the 'negative image set' contain no object in the concerned categories, which are used for exploring semi-supervised and weakly-supervised algorithms. The 650 images in the 'positive image set' were annotated with 10 categories of objects, which are used in the experiment and divided into a training set with 130 images, a validation set with 130 images, and a testing set with 390 images.

The DIOR data set is one of the most challenging large-scale benchmark data sets for RSI object detection. There are 23,463 images acquired from Google Earth, and 20 categories of 192,472 objects annotated in the DIOR data set. Compared with other data sets, the images and object instances of the data set have higher intra-class variation and inter-class similarity. Therefore, the DIOR data set is considered appropriate for the training and evaluation of RSI object detectors, especially deep-learning-based detectors. In the experiments, the quantities of the training set, the validation set, and the testing set are 5862, 5863, and 11,738, respectively, according to the official setting in [2].

3.2. Evaluation Metrics

In the experiments, the average precision (AP) for each category and mean average precision (mAP) are utilized to evaluate the proposed detectors. In general, the AP for the c-th category AP_c is calculated from recall values (R) and the corresponding precision values ($P_c(R)$) are calculated with formula (3), which is also the area under the precision–

recall curve of the category, and the mAP is calculated by averaging the AP_c over the C categories with formula (4).

$$AP_c = \int_0^1 P_c(R)dR \tag{3}$$

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP_c \tag{4}$$

For a specific category, to obtain the precision–recall curve, we need to calculate pairwise *Precision* values with formula (5) and *Recall* values with formula (6). Specifically, assume that there is a total of K bounding boxes classified into the category. Each prediction result includes coordinates and the classification confidence of a bounding box. The bounding box is true positive (TP) if the IOU between the ground-truth (GT) box and itself is larger than the threshold γ; otherwise, it is considered to be false positive (FP). In addition, if there is more than one TP bounding box corresponding to a GT box, the box with the largest IOU is reserved as TP, and the others are considered to be FP. If a GT box has no corresponding TP, then the GT box is considered false negative (FN). In formulas (5) and (6), the TP, FP, FN represent quantities of TP, FP, FN boxes; therefore, the *Precision* and *Recall* are dimensionless and $TP + FN$ is equal to the number of GT boxes $Num(GT)$. In practice, the bounding boxes are sorted according to their confidence, and the *Precision* and *Recall* values are calculated with the first k bounding boxes each time. The precision–recall curve is obtained by taking k from 1 to K. In the experiment, the IOU threshold γ is set to 0.5 according to the benchmarks of object detection in RSIs.

$$Precision = \frac{TP}{TP + FP} = \frac{TP}{K} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} = \frac{TP}{Num(GT)} \tag{6}$$

The *Precision* can be considered as the percentage of correct predictions out of all predictions, and the *Recall* can be the proportion of GT boxes that can be detected among all GT boxes. The precision–recall curve can reflect the relationship between *Precision* and *Recall*. A better detector should have both higher *Precision* and *Recall*, therefore its mAP should also be higher.

3.3. Baseline Methods

In the experiments, nine baseline methods, which are diffusely used as comparison benchmarks for object detection in RSIs, are adopted to evaluate the proposed detectors. To be specific, on the NWPU VHR-10 data set, the baseline methods include the traditional methods such as SSCBOW [5], and COPD [6], and deep-learning-based methods such as RICNN [10], R-P-Faster R-CNN [39], YOLO v3 [20], Deformable R-FCN [40], Faster RCNN [12], and Faster RCNN with FPN [17]. As for the DIOR data set, region-proposal-based methods including RICNN, Faster RCNN, Faster RCNN with FPN, and Mask RCNN [41] with FPN and the anchor-based method YOLO v3 are selected for a comprehensive comparison.

3.4. Implementation Details

ResNet [34] is recognized as one of the most effective backbone networks in the object-detection community. The residual operation of ResNet solves the degradation problem in deep networks; therefore, it can achieve a larger network and extract high-level semantic features. We adopt the ImageNet pre-trained ResNet-50 according to the choices of most baseline methods. To distinguish the feature maps of different scales, in addition to the 2D positional encoding, learnable scale-level encodings are also embedded in the multi-scale feature maps.

The encoder and decoder of the transformer both have six attention modules, and each module consists of eight attention heads. The dimension d of the input embeddings is

set to 256. The number of sampled keys for each deformable attention calculation K is set to 4. On the NWPU VHR-10 data set, the number of selected feature points N_p is set to 300. However, on the DIOR data set, the number is set to 600, because images may have more than 300 object instances in the DIOR data set.

The detectors are trained with the AdamW optimizer, setting the weight decay to 1×10^{-4}. The initial learning rate of the Transformer is set to 1×10^{-4}, while that of the other learnable parameters is set to 1×10^{-5}. The combined loss function in [28] is used for optimization, except that the section for classification is modified to the Focal Loss [42]. Other strategies of training and parameter initialization also follow [28].

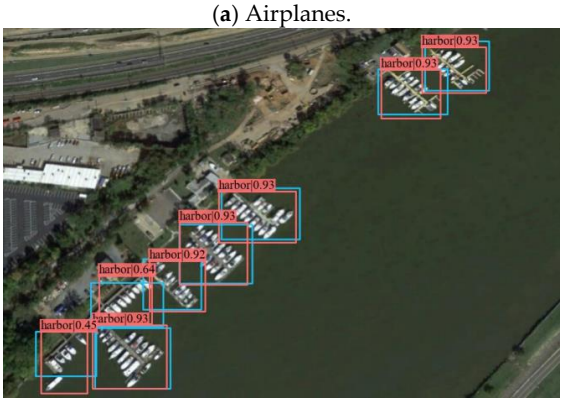
The proposed methods are implemented using MMDetection [43], which is an open-source object-detection framework presented by Open MMLab. The experiments are executed on a scientific computing workstation with Intel Xeon Silver CPUs and dual Tesla V100 MAX-Q GPUs with a total of 32 GB memory.

4. Experimental Results and Discussion

The proposed Transformer-based detectors are trained on the two data sets. Both qualitative inference results and quantitative evaluation results are provided and analyzed. For the qualitative inference results in Figures 6–8, the regions surrounded by blue bounding boxes indicate ground truth, and the detection results are marked with red bounding boxes. Additionally, the categories and confidence values of each detected box are given. In the quantitative evaluation results, the APs and mAPs magnified by 100 of the detectors are reported, and the precision–recall curve of each category is given. Additionally, the results of the ablation experiment are appended to provide the effectiveness of the modules in the proposed methods. For the quantitative evaluation results in Tables 1–5, the bold numbers represent the best performance compared to the other methods. At last, comparisons of the computational complexities and inference speeds between the proposed methods and baseline methods are exhibited.

(**a**) Airplanes.

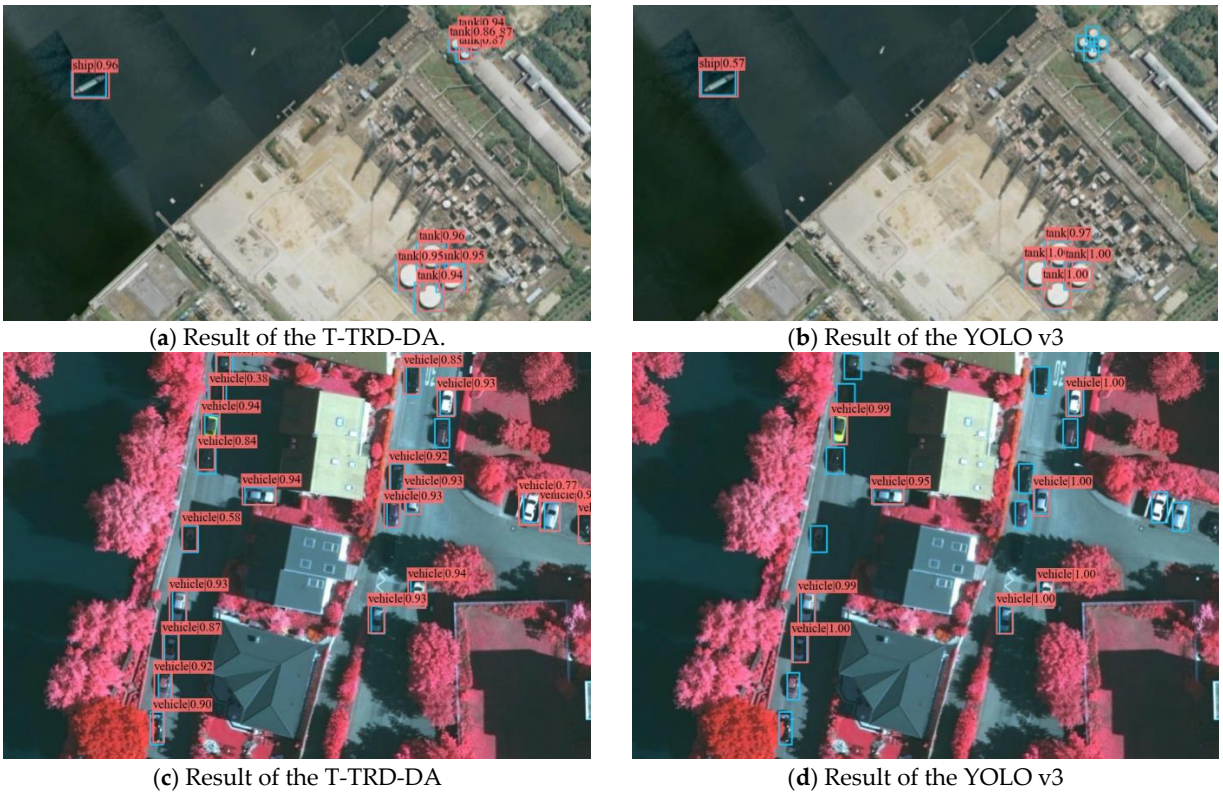
(**b**) Airplanes.



(**c**) Harbors.

(**d**) Baseball diamonds.

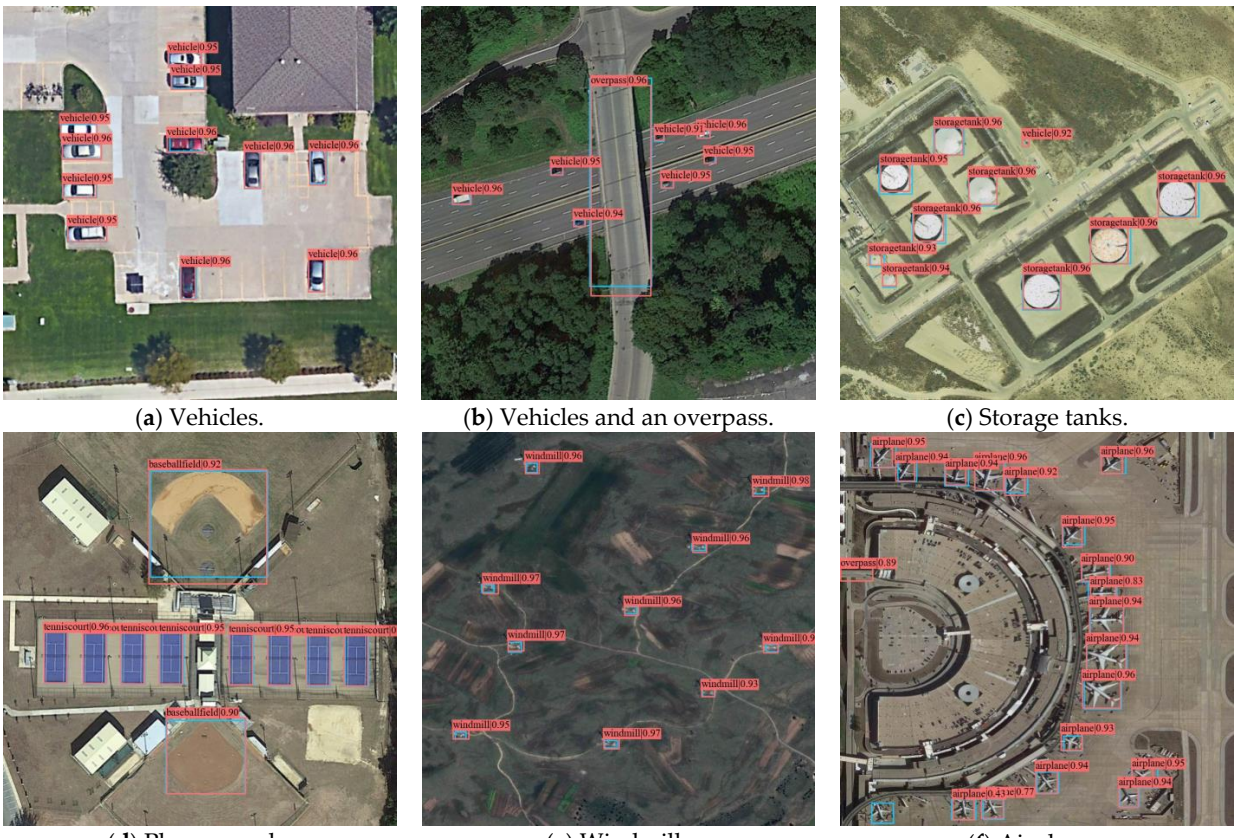
Figure 6. Qualitative inference results of T-TRD-DA on the NWPU VHR-10 data set.



(**a**) Result of the T-TRD-DA. (**b**) Result of the YOLO v3

(**c**) Result of the T-TRD-DA (**d**) Result of the YOLO v3

Figure 7. Comparison between T-TRD-DA (**left**) and YOLO v3 (**right**) on the NWPU VHR-10 data set.



(**a**) Vehicles. (**b**) Vehicles and an overpass. (**c**) Storage tanks.

(**d**) Play grounds. (**e**) Windmills (**f**) Airplanes.

Figure 8. Qualitative inference results on the DIOR data set.

Table 1. Comparison results of the proposed methods and baseline methods on the NWPU VHR-10 data set.

Method	AP (×100) for Each Category										mAP (×100)
	Plane	Ship	ST	BD	TC	BC	GTF	Harbor	Bridge	Vehicle	
SSCBOW [5]	50.6	50.8	33.4	43.5	00.3	15.0	10.1	58.3	12.5	33.6	30.8
COPD [6]	62.3	68.9	63.7	83.3	32.1	36.3	85.3	55.3	14.8	44.0	54.6
RICNN [10]	88.4	77.3	**85.3**	88.1	40.8	58.5	85.7	68.6	61.5	71.1	72.6
R-P-Faster R-CNN [39]	90.4	75.0	44.4	89.9	79.0	77.6	87.7	79.1	68.2	73.2	76.5
Yolo v3 [20]	90.6	63.1	70.9	94.8	83.8	68.6	**92.1**	76.2	58.1	65.7	76.4
Deformable R-FCN [40]	87.3	**81.4**	63.6	90.4	81.6	74.1	90.3	75.3	71.4	75.5	79.1
Faster RCNN [12]	92.0	76.0	54.1	95.4	75.6	71.3	90.1	76.0	69.0	63.8	76.3
Faster RCNN with FPN [17]	93.9	72.3	68.2	95.7	**91.9**	75.6	88.5	86.4	66.8	80.9	82.0
TRD	**99.4**	78.2	84.4	94.2	82.0	83.9	98.9	78.4	56.9	72.2	82.9
T-TRD-DA	99.0	81.0	79.6	**98.1**	89.2	**88.3**	86.5	**92.6**	**74.7**	**89.6**	**87.9**

Table 2. Results for objects of specific scale ranges on the NWPU VHR-10 data set.

Method	mAP (×100)	mAP (×100) for Objects of Different Scales		
		Large	Middle	Small
YOLO v3	76.4	74.2	69.9	52.3
Faster RCNN	76.3	76.5	73.0	35.2
Faster RCNN with FPN	82.0	77.4	79.5	47.9
TRD	82.9	79.8	75.6	43.7
T-TRD-DA	**87.9**	**80.8**	**83.6**	**65.7**

Table 3. Comparison results of the proposed methods and baseline methods on the DIOR data set.

Method		RICNN [10]	YOLO v3 [20]	Faster RCNN [12]	Faster RCNN with FPN [17]	Mask RCNN with FPN [41]	TRD	T-TRD-DA
AP (×100) for Each Class	Airplane	39.1	72.2	57.6	63.2	53.8	72.9	**77.9**
	Airport	61.0	29.2	68.6	61.3	72.3	79.3	**80.5**
	Baseball Field	60.1	**74.0**	62.4	66.3	63.2	70.0	70.1
	Basketball Court	66.3	78.6	83.7	85.5	81.0	83.8	**86.3**
	Bridge	25.3	31.2	31.2	36.0	38.7	38.8	**39.7**
	Chimney	63.3	69.7	73.9	73.9	72.6	77.8	**77.9**
	Dam	41.1	26.9	42.2	45.0	55.9	58.5	**59.3**
	ESA	51.7	48.6	55.0	56.9	**71.6**	57.6	59.0
	ETA	36.6	54.4	46.4	49.0	**67.0**	57.0	54.4
	Golf course	55.9	31.1	65.6	73.2	73.0	**75.2**	74.6
	Ground-track field	58.9	61.1	61.4	67.5	**75.8**	70.5	73.9
	Harbor	43.5	44.9	**52.2**	48.9	44.2	44.2	49.2
	Overpass	39.0	49.7	51.0	54.7	56.5	55.0	**57.8**
	Ship	9.1	**87.4**	48.0	73.2	71.9	73.5	74.2
	Stadium	61.1	**70.6**	51.0	62.8	58.6	52.1	61.1
	Storage Tank	19.1	68.7	35.3	68.3	53.6	67.6	**69.8**
	Tennis Court	63.5	**87.3**	73.5	78.7	81.1	82.5	84.0
	Train Station	46.1	29.4	50.3	51.4	54.0	56.0	**58.8**
	Vehicle	11.4	48.3	29.8	48.1	43.1	47.0	**50.5**
	Wind Mill	31.5	**78.7**	69.6	70.1	81.1	73.2	77.2
mAP (×100)		44.2	57.1	55.4	61.7	63.5	64.6	**66.8**

Table 4. Results for objects in specific scale ranges from the DIOR data set.

Method	mAP (×100)	mAP (×100) for Objects of Different Scales		
		Large	Middle	Small
Faster RCNN	55.4	80.7	43.7	7.9
Faster RCNN with FPN	61.7	81.9	45.7	18.4
TRD	64.6	83.2	50.1	20.4
T-TRD-DA	**66.8**	**93.1**	**67.2**	**33.3**

Table 5. Results of ablation experiment.

Method	mAP on NWPU VHR-10	mAP on DIOR
TRD	0.829	0.646
T-TRD	0.835	0.650
TRD-DA	0.866	0.664
T-TRD-DA	**0.879**	**0.668**

4.1. Comparison Results on the NWPU VHR-10 Data Set

Figure 6 shows the qualitative inference results of the proposed Transformer-based detectors on the NWPU VHR-10 data set. As illustrated in the figures, the proposed T-TRD-DA can detect most object instances in RSIs and correctly identify their categories. Even if the object instances are small, which are hard to detect, the T-TRD-DA still performs well. Figure 7 provides a qualitative comparison between the proposed T-TRD-DA and YOLO v3. In Figure 7a,b, the smaller storage tanks are all detected by the proposed T-TRD-DA, while the YOLO v3 omits some of them. In Figure 7c,d, the T-TRD-DA recognizes almost all vehicles, while the YOLO v3 leaves out more than half of them. As a consequence, in contrast to YOLO v3, the proposed T-TRD-DA is shown not to be susceptible to objects of small scale, clustered objects, or objects being obscured by shallows, etc.

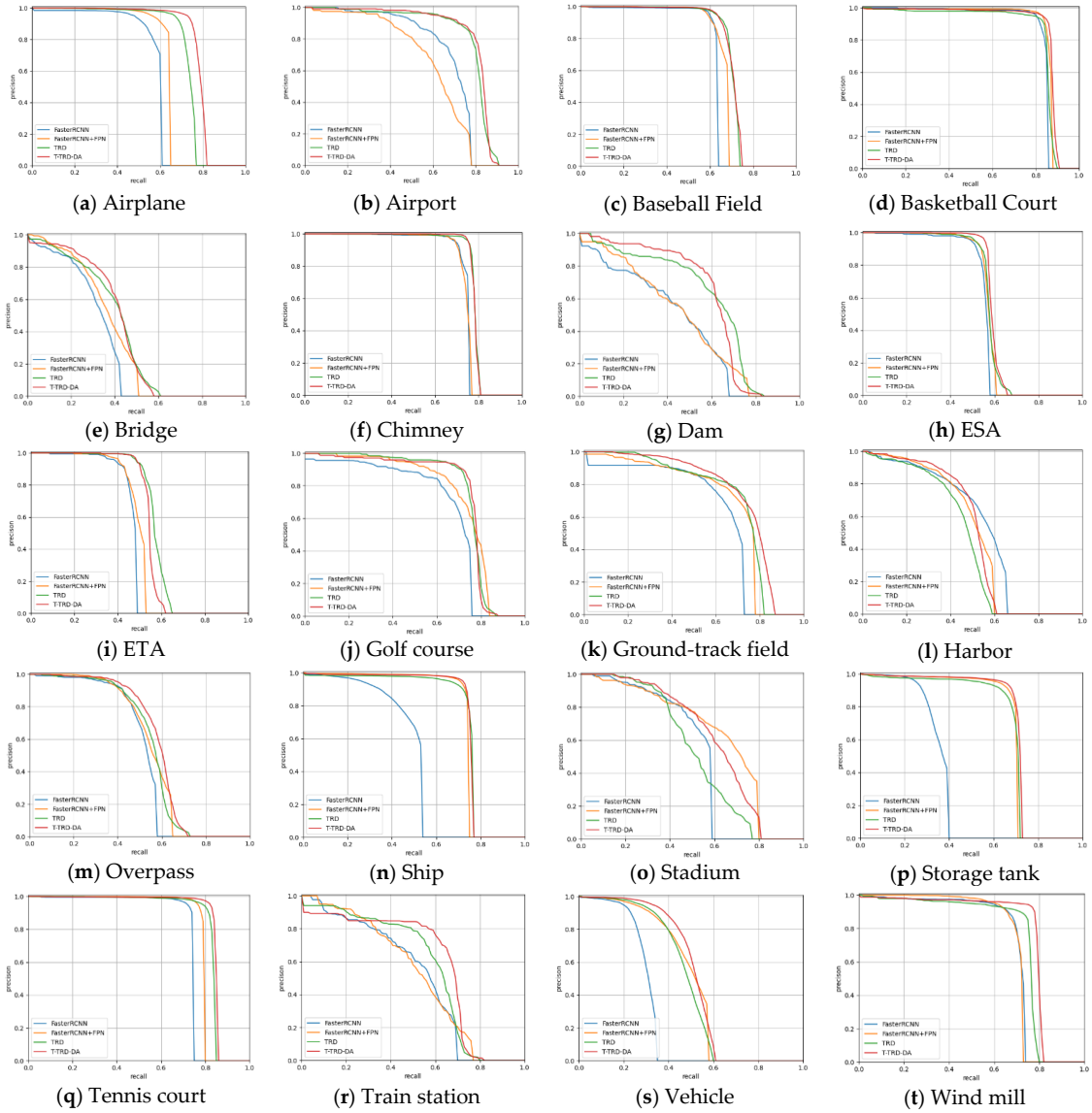
Table 1 shows the comparison results on the NWPU VHR-10 data set, where ST denotes the storage tank, BD denotes the baseball diamond, TC denotes the tennis court, BC denotes the basketball court, and GT denotes the ground-track field. As shown in the table, the CNN-based methods exhibit a noticeable advantage compared to the traditional BOW-based SSCBOW method and the SVM-based COPD method. Among these CNN-based methods for object detection in RSIs, the Faster RCNN is the most representative one, which can swiftly provide region proposals and then make precise predictions. The FPN is often used for the multi-scale feature fusion of features extracted from the CNN backbone, which effectively enhances the detection capability of small-object instances. Therefore, the Faster RCNN with FPN is a relatively competitive baseline method for object detection in RSIs. Nevertheless, the proposed TRD outperforms all the baseline methods and surpasses the Faster RCNN with FPN baseline with a 0.02 improvement on the mAP. With the same backbone to extract features of RSIs, the Transformer-based detection head of the TRD exhibits its powerful detection capability and exceeds the CNN-based detection heads, which demonstrates the feasibility of using the Transformer for object detection in RSIs. Furthermore, with the promotion of the proposed attention-based transferring backbone and data augmentation, the T-TRD-DA achieves a better detection performance, with an mAP that reaches 0.879 and obtains outstanding APs in all categories. As a consequence, the improvements can make efficient progress on the proposed Transformer-based RSI object-detection framework.

Additionally, the comparison results of the proposed methods and baseline methods on objects of specific scale ranges, i.e., large, middle, small, are reported in Table 2. The mAP of the Faster RCNN baseline is limited in its detection of small objects because its backbone only outputs the highest-level features, which have low resolution and cause poor detection performance. The FPN capable of multi-scale feature fusion effectively solves this problem. Therefore, the Faster RCNN with FPN baseline achieves great improvement on small

objects. The proposed TRD and T-TRD-DA can aggregate multi-scale features without FPN, and they also have outstanding detection capacities for small objects. Moreover, the proposed Transformer-based detectors also perform well on large objects and middle objects, which means a better overall detection capability.

4.2. Comparison Results on the DIOR Data Set

To further evaluate the effectiveness of the proposed Transformer-based detectors, the detectors are trained on the DIOR data set and compared with more competitive baseline methods. Figure 8 shows the qualitative inference results of the proposed T-TRD-DA on the DIOR data set. It is obvious that the proposed T-TRD-DA exhibits an intuitively satisfactory detection capability on the large-scale challenge data set. The precision–recall curves of each category are provided in Figure 9, which intuitively shows the detailed relationship between precision and recall. The ETA and ESA are the abbreviations of expressway toll station and expressway service area, respectively. It can be seen that the proposed T-TRD-DA detector exhibits a superior performance in most categories, such as airplane, ground-track field, tennis court, etc.



(**a**) Airplane (**b**) Airport (**c**) Baseball Field (**d**) Basketball Court

(**e**) Bridge (**f**) Chimney (**g**) Dam (**h**) ESA

(**i**) ETA (**j**) Golf course (**k**) Ground-track field (**l**) Harbor

(**m**) Overpass (**n**) Ship (**o**) Stadium (**p**) Storage tank

(**q**) Tennis court (**r**) Train station (**s**) Vehicle (**t**) Wind mill

Figure 9. The p–r curve of the detectors on each category of the DIOR data set.

Table 3 shows the results of the DIOR data set and compares the proposed TRD and T-TRD-DA to five representative deep-learning-based methods, including the AP values of the 20 categories and the mAP. In these baseline methods, the Mask RCNN, which was originally designed for object-instance segmentation, is extended from the Faster RCNN and achieves state-of-the-art performance of object detection. With the FPN, both the Faster RCNN and Mask RCNN can detect objects with a wide variety of scales and acquire great advances to their overall detection performance. Additionally, as shown in Table 4, compared to the Faster RCNN and the Faster RCNN with FPN, the proposed TRD acquires outstanding detection capacity for the three scale ranges, especially on the small objects. The proposed T-TRD-DA achieves the best performance, which is attributed to the multi-scale-feature embedding. Above all, with the powerful context-modeling capabilities of the Transformer, the proposed Transformer-based detectors can accurately detect the objects of interest in the complicated RSIs.

4.3. Ablation Experiments

Four sets of ablation experiments on both data sets are performed to evaluate the efficiencies of the improvements to the proposed T-TRD-DA, and the results are reported in Table 5. The results indicate that both the improvements to the attention-based transferring backbone and the data augmentation benefit the detection performance of the TRD. The transferring backbone utilizes the knowledge learned from the source-domain data to extract more effective features of the RSIs, and then uses the attention mechanism to adaptively regulate the channel-wise features. Additionally, the data augmentation enriches the orientations, scales, and backgrounds of the object instances, which strengthens the generalization performance of the detectors. Therefore, the final T-TRD-DA achieves a competitive detection capability and indicates the great potential of the Transformer for RSI object detection.

4.4. Comparison of the Computational Complexity and Inference Speed

To evaluate the computational efficiency of the methods, the values of floating-point operations (FLOPs) and the inference speeds of the proposed Transformer-based methods and three baseline methods are reported in Table 6. The FLOPs and FPS of each method are measured with the analysis tools of MMDetection, with inputs of 800 × 800 size from both data sets. As is shown, the FLOPs of the proposed Transformer-based-detection models are close to the models of the baseline methods, and are only higher than YOLO v3. However, due to the high computational cost of the Transformer, the inference speeds are still able to be improved.

Table 6. Comparison of the computational complexity and inference speed.

Method	NWPU VHR-10		DIOR	
	FLOPs (G)	Inference FPS	FLOPs (G)	Inference FPS
YOLO v3	121.27	42.8	121.41	33.2
Faster RCNN	127.91	27.5	127.93	26.3
Faster RCNN with FPN	135.25	22.4	135.30	19.6
TRD	125.63	14.2	125.67	13.2
T-TRD-DA	125.70	13.2	125.74	12.5

4.5. Discussion

In the experiments, the proposed Transformer-based methods were evaluated and compared with the state-of-the-art CNN-based RSI object-detection frameworks. The experiments demonstrated the effectiveness of the proposed Transformer-based frameworks and their advantages over the CNN-based frameworks.

From the qualitative inference results in Figures 6–8, it could be seen that the proposed T-TRD-DA could accurately recognize objects of various categories, scales, and orientations.

The bounding boxes of prediction were highly closed to the GT boxes. Additionally, from the quantitative evaluation results in Tables 1 and 3, the TRD and T-TRD-DA achieved 82.9 and 87.9 on the NWPU VHR-10 data set, and obtained 64.6 and 66.8 on the DIOR data set in terms of centuple mAP, respectively.

From the ablation experiments in Table 5, compared with the TRD, the proposed T-TRD obtained an improvement of 0.6 in terms of centuple mAP on the NWPU VHR-10 data set. It was not a great success, but it showed that the proper adjustment of the feature map led to better RSI detection performance. Moreover, the TRD-DA improved by 3.7 in terms of centuple mAP on the NWPU VHR-10 data set. The overfitting problem caused by limited training samples was mitigated by the data augmentation in the TRD-DA. With the two improvements, the proposed T-TRD-DA improved by 5.0 in terms of centuple mAP on the NWPU VHR-10 data set. Therefore, the attention-based transferring backbone and the data augmentation were both efficient and indispensable in the proposed T-TRD-DA.

From Tables 1 and 3, the proposed TRD and T-TRD-DA methods both exceeded all the competitive CNN-based RSI object-detection methods. For example, Faster RCNN only obtained 0.554 in terms of mAP on the DIOR dataset. The proposed TRD, which was based on a well-designed Transformer, obtained 0.646 in terms of mAP on the DIOR dataset. The results of the comparison experiments revealed the advantages of the proposed Transformer-based methods, which are discussed as follows.

Firstly, CNN-based methods were good at object detection. However, for RSI object-detection tasks, due to the large spatial size (e.g., the spatial size of DIOR data set is 800×800), it was difficult to obtain the global representation of RSIs. The Transformer was good at capturing long-distance relationships, hence it could obtain more discriminative features.

Secondly, CNN-based methods usually required FPN [14] for multi-scale feature fusion to improve the performance on small objects. From Tables 2 and 4, the TRD and T-TRD-DA performed better on objects of various scales than the CNN-based methods with FPN, especially on small objects. In contrast to the FPN, which added the down-sampled features at the same positions of all the scales, the proposed Transformer-based frameworks could adaptively integrate features at various crucial positions of different scales; therefore, it achieved impressive small-object-detection performance.

Additionally, the representative CNN-based frameworks, such as the Faster RCNN [12] or YOLO v3 [20], were usually based on anchors. However, the setting of sizes, amount, and aspect ratios for anchor generation affected the detection performances. The proposed TRD and T-TRD-DA aggregated the pyramidal features and acquired spatial- and level-aware feature points for representing instances. Therefore, the proposed methods were anchor-free and convenient to train.

Moreover, from Table 6, although deformable attention was developed in the TRD and T-TRD-DA to simplify the calculation of the Transformer, the inference speeds of the proposed methods were slower than those of the CNN-based methods. More research into the improvement of inference speed is required.

Above all, in this study, a modified Transformer combined with a transfer CNN was proposed for RSI object detection. Elaborate experiments and analyses have indicated the superiorities of the proposed Transformer-based frameworks. Besides, the disadvantages have also been analyzed for further research on developing Transformer-based RSI object-detection methods.

5. Conclusions

In this study, Transformer-based frameworks were explored for RSI object detection. It was found that the Transformer was good at obtaining the long-distance relationship; therefore, it could capture global spatial- and scale-aware features of RSIs and detect objects of interest. The proposed TRD used the pre-trained CNN to extract local discriminate features, and the Transformer was modified to process feature pyramid of an RSI and predict the categories and the box coordinates of the objects in an end-to-end manner. By combining the advantages of the CNN and Transformer, the experimental results of diverse

terms demonstrated that the TRD achieved impressive RSI object-detection performance for objects of different scales, especially on small objects.

There was still a lot of room for improvement in the TRD. On the one hand, the use of the pre-trained CNN faced the problem of data-set shift (i.e., the source data set and the target data set were quite different). On the other hand, there were insufficient training samples for RSI object detection to train a Transformer-based model. Hence, to further improve the performance of the TRD, an attention-based transferring backbone and data augmentation were combined with the TRD to formulate the T-TRD-DA. The ablation experiments on various structures, i.e., TRD, T-TRD, TRD-DA, and T-TRD-DA, have shown that the two improvements as well as their combination were efficient. The T-TRD-DA was proved to be a state-of-the-art RSI object-detection framework.

Compared with the CNN-based frameworks, the proposed T-TRD-DA was demonstrated to be a better detection architecture. There were not anchors, non-maximum suppression, or FPN in the proposed frameworks. However, the T-TRD-DA exceeded YOLO-v3 and the Faster RCNN with FPN in detecting small objects. As an early stage of the Transformer-based detection method, the T-TRD-DA showed the potential of the Transformer-based RSI object-detection methods. Nevertheless, the proposed Transformer-based frameworks have the problem of low inference speed, which is another topic for further research.

Very recently, some modifications of the Transformer, including the self-training Transformer and transferring Transformer, can be investigated for RSI object detection in the near future.

The findings reported in this study have some implications for effective RSI object detection, which show that Transformer-based methods have huge research value in the area of RSI object detection.

Author Contributions: Conceptualization: Y.C.; methodology: Q.L. and Y.Z.; writing—original draft preparation: Q.L. and Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China under the Grant 61971164 and the Grant U20B2041.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experiments are evaluated on publicly open data sets. The access manner of the data sets can refer to the corresponding published papers.

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

References

1. Cheng, G.; Han, J. A survey on object detection in optical remote sensing images. *ISPRS J. Photogramm. Remote Sens.* **2016**, *117*, 11–28. [CrossRef]
2. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object detection in optical remote sensing images: A survey and a new benchmark. *ISPRS J. Photogramm. Remote Sens.* **2019**, *159*, 296–307. [CrossRef]
3. Lou, X.; Huang, D.; Fan, L.; Xu, A. An image classification algorithm based on bag of visual words and multi-kernel learning. *J. Multimed.* **2014**, *9*, 269–277. [CrossRef]
4. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
5. Sun, H.; Sun, X.; Wang, H.; Li, Y.; Li, X. Automatic target detection in high-resolution remote sensing images using spatial sparse coding bag-of-words model. *IEEE Geosci. Remote Sens. Lett.* **2012**, *9*, 109–113. [CrossRef]
6. Cheng, G.; Han, J.; Zhou, P.; Guo, L. Multi-class geospatial object detection and geographic image classification based on collection of part detectors. *ISPRS J. Photogramm. Remote Sens.* **2014**, *98*, 119–132. [CrossRef]
7. Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012.
8. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.

9. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13–16 December 2015.
10. Cheng, G.; Zhou, P.; Han, J. Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2016**, *54*, 7405–7415. [CrossRef]
11. Cheng, G.; Han, J.; Zhou, P.; Xu, D. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. *IEEE Trans. Image Process.* **2019**, *28*, 265–278. [CrossRef]
12. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
13. Li, K.; Cheng, G.; Bu, S.; You, X. Rotation-insensitive and context-augmented object detection in remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 2337–2348. [CrossRef]
14. Lin, T.Y.; Dollar, 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.
15. Zhang, X.; Zhu, K.; Chen, G.; Tan, X.; Zhang, L.; Dai, F.; Liao, P.; Gong, Y. Geospatial object detection on high resolution remote sensing imagery based on double multi-scale feature pyramid network. *Remote Sens.* **2019**, *11*, 755. [CrossRef]
16. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-scale object detection in remote sensing imagery with convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2018**, *145*, 3–22. [CrossRef]
17. Guo, W.; Yang, W.; Zhang, H.; Hua, G. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. *Remote Sens.* **2018**, *10*, 131. [CrossRef]
18. Redmon, J.; Divvala, S.; 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.
19. 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.
20. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
21. Pham, M.-T.; Courtrai, L.; Friguet, C.; Lefèvre, S.; Baussard, A. YOLO-Fine: One-Stage Detector of Small Objects Under Various Backgrounds in Remote Sensing Images. *Remote Sens.* **2020**, *12*, 2501. [CrossRef]
22. 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]
23. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Las Vegas, NV, USA, 27–30 June 2016.
24. Zhuang, S.; Wang, P.; Jiang, B.; Wang, G.; Wang, C. A Single Shot Framework with Multi-Scale Feature Fusion for Geospatial Object Detection. *Remote Sens.* **2019**, *11*, 594. [CrossRef]
25. 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 2021.
26. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Gelly, S. An image is worth 16 $\times$ 16 words: Transformers for image recognition at scale. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021.
27. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv* **2021**, arXiv:2103.14030.
28. Nicolas, C.; Francisco, M.; Gabriel, S.; Nicolas, U.; Alexander, K.; Sergey, Z. End-to-End Object Detection with Transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020.
29. He, X.; Chen, Y.; Lin, Z. Spatial-Spectral Transformer for Hyperspectral Image Classification. *Remote Sens.* **2021**, *13*, 498. [CrossRef]
30. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking hyperspectral image classification with transformers. *arXiv* **2021**, arXiv:2107.02988. [CrossRef]
31. Zhang, J.; Zhao, H.; Li, J. TRS: Transformers for Remote Sensing Scene Classification. *Remote Sens.* **2021**, *13*, 4143. [CrossRef]
32. Zheng, Y.; Sun, P.; Zhou, Z.; Xu, W.; Ren, Q. ADT-Det: Adaptive Dynamic Refined Single-Stage Transformer Detector for Arbitrary-Oriented Object Detection in Satellite Optical Imagery. *Remote Sens.* **2021**, *13*, 2623. [CrossRef]
33. 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]
34. 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, 27–30 June 2016.
35. Oquab, M.; Bottou, L.; Laptev, I.; Josef, S. Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
36. Lin, Z.; Feng, M.; Santos, C.N.D.; Yu, M.; Xiang, B.; Zhou, B.; Bengio, Y. A structured self-attentive sentence embedding. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
37. Aurelio, Y.; Almeida, G.; Castro, C.; Braga, A. Learning from imbalanced data sets with weighted cross-entropy function. *Neural Process. Lett.* **2019**, *50*, 1937–1949. [CrossRef]
38. Michael, C. The DGPF-test on digital airborne camera evaluation overview and test design. *PFG Photogramm.-Fernerkund. Geoinf.* **2010**, *2*, 73–82.

39. Han, X.; Zhong, Y.; Zhang, L. An efficient and robust integrated geospatial object detection framework for high spatial resolution remote sensing imagery. *Remote Sens.* **2017**, *9*, 666. [CrossRef]
40. Xu, Z.; Xu, X.; Wang, L.; Yang, R.; Pu, F. Deformable ConvNet with aspect ratio constrained NMS for object detection in remote sensing imagery. *Remote Sens.* **2017**, *9*, 1312. [CrossRef]
41. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
42. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
43. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv* **2019**, arXiv:1906.07155.