remote sensing

MDPI

Article

DHQ-DETR: Distributed and High-Quality Object Query for Enhanced Dense Detection in Remote Sensing

Chenglong Li, Jianwei Zhang *, Bihan Huo and Yingjian Xue

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing 210044, China; 202212150015@nuist.edu.cn (C.L.); lizimo@nuist.edu.cn (B.H.); 202312150037@nuist.edu.cn (Y.X.)

* Correspondence: zhangjw@nuist.edu.cn

Abstract: With the widespread application of remote sensing technologies and UAV imagery in various fields, dense object detection has become a significant and challenging task in computer vision research. Existing end-to-end detection models, particularly those based on DETR, often face criticism due to their high computational demands, slow convergence rates, and inadequacy in managing dense, multi-scale objects. These challenges are especially acute in remote sensing applications, where accurate analysis of large-scale aerial and satellite imagery relies heavily on effective dense object detection. In this paper, we propose the DHQ-DETR framework, which addresses these issues by modeling bounding box offsets as distributions. DHQ-DETR incorporates the Distribution Focus Loss (DFL) to enhance residual learning, and introduces a High-Quality Query Selection (HQQS) module to effectively balance classification and regression tasks. Additionally, we propose an auxiliary detection head and a sample assignment strategy that complements the Hungarian algorithm to accelerate convergence. Our experimental results demonstrate the superior performance of DHQ-DETR, achieving an average precision (AP) of 53.7% on the COCO val2017 dataset, 54.3% on the DOTAv1.0, and 32.4% on Visdrone, underscoring its effectiveness for real-world dense object detection tasks.

Keywords: detection transformer; dense object detection; multi-scale object detection; distribution modeling

Academic Editor: Wen Yang

Received: 20 January 2025
Revised: 28 January 2025
Accepted: 30 January 2025
Published: 1 February 2025

Citation: Li, C.; Zhang, J.; Huo, B.; Xue, Y. DHQ-DETR: Distributed and High-Quality Object Query for Enhanced Dense Detection in Remote Sensing. *Remote Sens.* **2025**, *17*, 514. https://doi.org/10.3390/rs17030514

1. Introduction

Object detection is a foundational element in computer vision, involving the determination of object locations and their categorical classification. Its significance transcends general applications, particularly in specialized fields such as remote sensing and Unmanned Aerial Vehicle (UAV) imagery. In these areas, object detection is crucial for tasks including monitoring environmental changes, identifying urban structures, and evaluating disaster zones. Dense object detection is particularly important in these contexts due to the high density and complexity of objects in satellite and aerial imagery, necessitating precise mapping and real-time decision-making.

Presently, the architectures employed in object detection can be broadly divided into two categories: convolutional neural networks (CNNs) and Transformers. The development and refinement of CNN-based object detectors have been the subject of extensive research. Initially, a two-stage strategy predominant in the field involved region proposal followed by classification. However, these approaches often suffer from performance degradation when dealing with low-resolution imagery, and their intricate pipeline hampers real-time inference capabilities [1–3]. This shortcoming has paved the way for efficient one-stage

detectors, which have emerged as a significant area of focus in both research and application. These detectors generate dense predictions directly on feature maps, thus achieving a more favorable trade-off between accuracy and computational efficiency [4–11]. Some approaches utilize anchor boxes, adjusted by offsets and scales derived from prior knowledge [6–10], while others advocate for anchor-free methods, emphasizing greater adaptability and criticizing the lack of versatility in anchor-based designs [4,5,12]. However, one-stage detectors are not without limitations, and overcoming these presents a significant challenge for further advancements. For instance, these detectors find it challenging to make sparse predictions without compromising on detection performance and recall rates. Additionally, the dense predictions they generate do not naturally align with the desired outcome, necessitating manual post-processing steps, such as non-maximum suppression (NMS), to eliminate duplicate predictions. This additional step introduces delays in inference, and may degrade accuracy. While improved NMS methods have been proposed, their generalizability remains limited. Moreover, the literature has yet to adequately address challenges surrounding dense multi-scale object detection [13–15].

The Detection Transformer (DETR) has transformed the object detection paradigm by recasting it as a set prediction problem, leveraging transformer encoder–decoder architectures and the Hungarian algorithm for matching [16]. As a pioneering method, DETR has initiated a new direction in object detection research, yet there is ample room for refinement in areas such as training methodologies and computational efficiency. For example, Deformable DETR improves upon DETR's slow convergence and performance on small objects by incorporating multi-scale deformable attention modules [17]. DAB-DETR enhances the interpretability of DETR by integrating anchor boxes into the decoder's query modeling [18]. DN-DETR accelerates convergence by bypassing the Hungarian matching process, instead feeding noisy ground truth directly to the decoder [19]. Group DETR leverages one-to-many training relationships by introducing multiple object queries, thereby improving performance [20]. Sparse-DETR reduces computational demands in the encoder through the use of sparse queries, while Focus-DETR achieves a refined balance between model complexity and accuracy by selectively focusing on 30% of the foreground tokens and semantic features [21]. RT-DETR further reduces computation without sacrificing performance by restricting multi-head self-attention to downsampled features at a factor of 32 [22].

The Detection Transformer (DETR) has been the subject of extensive research, exploring various methods to enhance its performance. Nonetheless, DETR continues to encounter challenges, particularly in the realm of dense multi-scale object detection, where its performance remains limited. To address these challenges, we introduce DHQ-DETR, a novel approach that redefines the paradigm of object detection. Firstly, we revolutionize the concept of the bounding box by positing that the ground truth adheres to a Dirac distribution. In conjunction with the Intersection over Union (IoU) loss, we implement the distribution focus loss (DFL), which exhibits consistent performance in managing objects across diverse scales. Secondly, we integrate a high-quality query selection module to improve the initialization of object queries, thereby ensuring better alignment between classification and regression tasks. Lastly, drawing inspiration from CO-DETR, we introduce an additional detection head and a refined assignment method that increases the number of positive samples in the decoding layer. This enhancement not only boosts the stability of the Hungarian algorithm, but also accelerates model convergence [23,24].

In this paper, we unveil DHQ-DETR, an end-to-end object detection model, the architecture of which is depicted in Figure 1. DHQ-DETR achieves notable detection results, including an AP of 53.7% on the COCO val2017 dataset, an AP of 54.3% on the DOTAv1.0

test set, and an AP of 32.4% on the Visdrone test set. The principal contributions of our research are summarized as follows:

1. We propose a groundbreaking distribution-based approach to box modeling and incorporate the distribution focus loss, which demonstrates robustness when dealing with dense multi-scale targets.
2. We introduce a high-quality query selection module designed to resolve the misalignment inherent in the initialization of object queries.
3. We develop a refined assignment strategy, coupled with an extra detection head, to enhance the stability and convergence speed of the DETR training process.

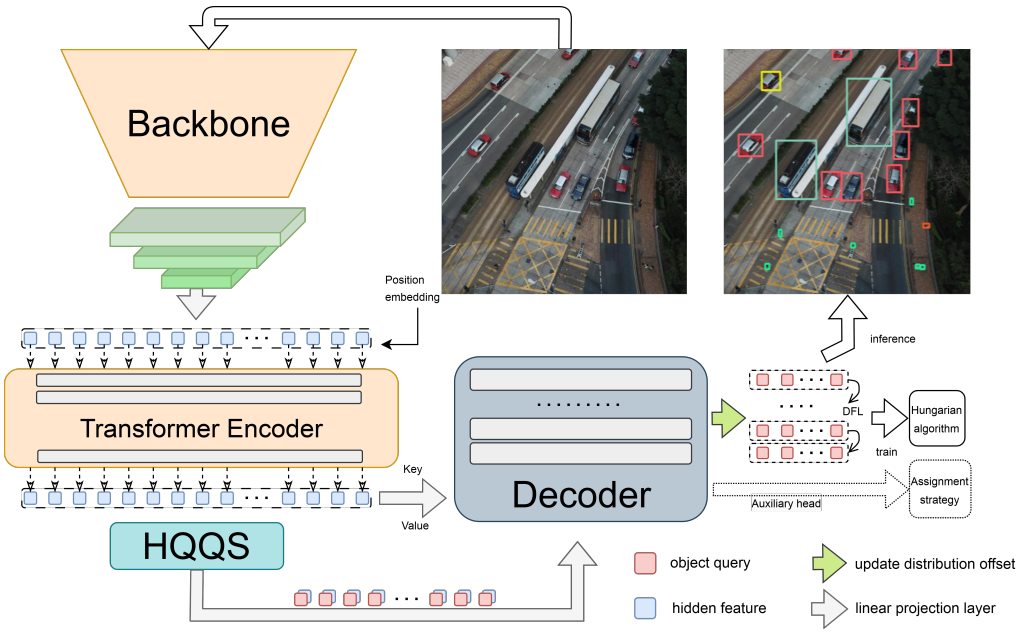

Figure 1. Illustration of the Distribution-DETR architecture. The High-Quality Query Selection (HQQS) module is employed to initialize object queries aligned with both classification and localization tasks. By modeling the offsets between predictions and labels as distributions, the model enhances its capacity to localize occluded target boundaries. Furthermore, the decoder integrates an assignment strategy to strengthen positive supervision.

The remainder of this paper is structured as follows: Section 2 provides a brief review of the relevant background work and theoretical foundations, with a focus on the development of DETR and its variants. Section 3 details the proposed DHQ-DETR model, highlighting its key components and innovative training strategies. Section 4 outlines the experimental setup, datasets, evaluation metrics, as well as a comparative analysis with baseline models. Finally, Section 5 summarizes our research and discusses potential future research directions.

2. Related Work

2.1. CNN-Based Detectors

CNN-based detectors frequently employ pyramid feature representations, a cornerstone technique in the field of object detection. Initially introduced in SSD, these pyramids are designed for the efficient detection of both small and large-scale targets by leveraging high-resolution feature maps for the former and low-resolution maps for the latter [8]. To integrate a richer semantic spectrum and capture multi-scale features, components such as FPN and PAN have been developed to meld features between the backbone network and the detection head [25,26]. Furthermore, incorporating plug-and-play attention modules, including SENet and CBAM, has proven effective in tackling multi-scale detection

challenges by applying attention weights in either the spatial or feature dimensions [27,28]. These modules help overcome the intrinsic limitations of convolutional neural networks and enhance the model's ability to represent features effectively. CNN-based detectors require post-processing to eliminate redundant predictions. Numerous studies have investigated the optimization of post-processing methods. For example, Soft-NMS mitigates the impact of highly overlapping boxes by decreasing their scores rather than removing them outright [13]. Adaptive NMS modifies the suppression threshold dynamically, contingent on the distribution of various object classes within overlapping regions [15]. ConvNMS replaces the traditional threshold computation with convolutional kernels [29]. Learning NMS involves training the model to independently acquire appropriate suppression strategies [30]. WBF determines the ultimate fused bounding box position and dimensions through weighted voting, based on the overlap and confidence of the bounding boxes [31]. Additionally, there is a push towards employing more adaptable techniques for defining bounding boxes. Gaussian YOLOv3, for instance, incorporates Gaussian modeling to refine the location information of the bounding box, incorporating mean and variance to quantify the uncertainty of the location, which is then integrated into NMS [32,33]. However, the assumption that bounding boxes adhere to a Gaussian distribution is simplistic, as their actual distribution tends to be more arbitrary and flexible. Consequently, Generalized Focal Loss (GFL) explores the use of arbitrary distributions in depicting bounding boxes within convolution-based detectors, which also informs the present research [34].

2.2. End-to-End Object Detector

In contrast to CNN-based detectors, the DETR model obviates the need for post-processing, thanks to its global self-attention architecture, which shows promise in detecting large-scale or occluded targets [16]. However, the full potential of DETR has yet to be realized compared to CNN-based detectors. This has led to the proposal of various enhanced and generalized methods for DETR. For instance, Deformable DETR improves convergence speed and performance on small objects by integrating a Multi-Scale Deformable Attention Module [17]. DAB-DETR models the decoder queries of DETR using four-dimensional anchor boxes [18]. DN-DETR fed noisy ground truth boxes directly to the decoder, learning relative offsets through shortcut connections [19]. Group DETR introduces multiple object queries and leverages a one-to-many advantage during training to enhance detection performance [20]. Sparse-DETR reduces the computational complexity in the encoder by utilizing sparse queries [21]. Focus-DETR achieves comparable performance by selecting approximately 30% of foreground tokens and semantically enriching fine-grained features [35]. RT-DETR restricts global multi-head self-attention to features downsampled by 32 times, diminishing computational expense without degradation in model performance [22]. Despite these improvements, the majority focus on enhancing training methods, convergence speed, or computational efficiency, rather than addressing dense multi-scale detection. The adoption of DETR in remote sensing faces significant challenges due to the pronounced scale variations between pedestrians and vehicles in UAV aerial images from the VisDrone dataset or DOTAv1.0, as well as the occlusion challenges in crowded scenes, as illustrated in Figure 2 [36].

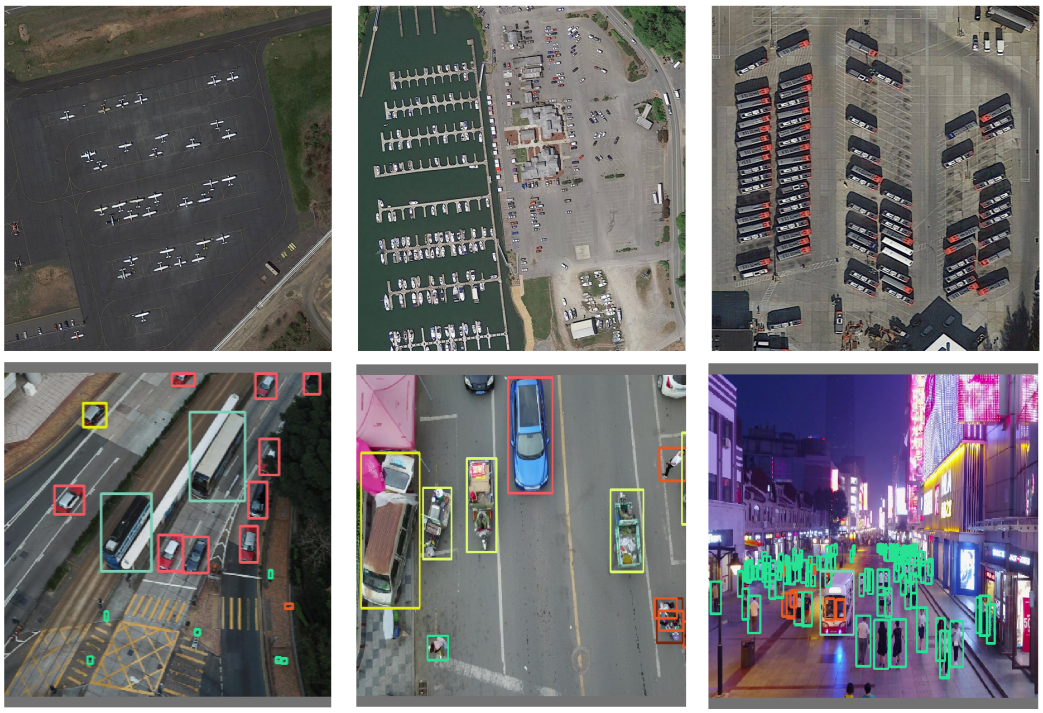

Figure 2. The figure highlights the challenges in remote sensing target detection, where variations in density and scale complicate detection. Targets may be tightly clustered or vary greatly in size, making accurate identification difficult, especially in aerial or satellite imagery.

3. Materials and Methods

In Section 3.1, we present the overall structure of the DHQ-DETR model using a system architecture diagram, highlighting the main components and their interactions. Section 3.2 provides a detailed account of the distribution offset estimation method and describes the design of the distribution focus loss (DFL), while Section 3.3 explains the implementation of the high-quality query selection (HQQS) module. In Section 3.4 we detail the development and function of the short-circuit training decoder and auxiliary detection head. This organization aims to give readers a clearer understanding of the model's components and their collaboration to enhance object detection performance.

3.1. The Overall Structure

The overall framework of DHQ-DETR is illustrated in Figure 1. Unlike DETR, DHQ-DETR employs a high-quality query selection module to diversify object query initialization, thereby reducing redundant high-confidence features. Furthermore, in the decoder, instead of directly predicting the offset of the anchor box in the Sigmoid domain, the model predicts the probability distribution of predefined offset values and integrates these predictions to determine the updated anchor box value. Additionally, across decoding layers, Distribution Focal Loss (DFL) guides the distribution towards convergence to a Dirac distribution. Finally, to enhance the convergence speed of DETR, we incorporate an auxiliary prediction head and propose a novel positive and negative sample allocation method that aligns classification and regression tasks. Supplementing the Hungarian algorithm, this approach significantly improves the stability and convergence speed of model training.

3.2. Distribution-Based Modeling

3.2.1. Basic Decoder

The decoding scheme presented in Figure 3a is prevalent in most contemporary DETR-like models. It comprises several decoding layers tasked with adjusting anchor positions.

Initially, the cross-attention mechanism in the decoding layer $decoder_{i+1}$ updates the hidden features F_i^{hide}, as depicted in the following equation:

$$F_{i+1}^{hide} = decoder_{i+1}(F_i^{hide}, Q_i, K, V), \tag{1}$$

where the variable Q_i represents the object query derived from the preceding layer. The decoding layer $decoder_{i+1}$ utilizes Q_i to extract local features, which subsequently inform the update of the hidden features F_i^{hide}. The key and value, denoted by K and V, respectively, are obtained from the outputs of the encoder. Following this, a feedforward network FFN_{dec} is employed to compute four offsets, Δx_{i+1}, corresponding to adjustments in the center coordinates, length, and width. FFN is a module composed of several fully connected and activated layers, which maps the features of the latent space into offsets of the four dimensions of the anchor box. The parameters of FFN_{dec} are shared across multiple decoding layers, which allows for the following representation:

$$\Delta x_{i+1} = FFN_{dec}(F_{i+1}^{hide}). \tag{2}$$

The computed offset is then applied to the object query to finalize the adjustments as follows:

$$Q_{i+1} = \sigma(\Delta x_{i+1} + \widehat{\sigma}(Q_i)), \tag{3}$$

where σ and $\widehat{\sigma}$ denote the sigmoid and inverse sigmoid functions, respectively.

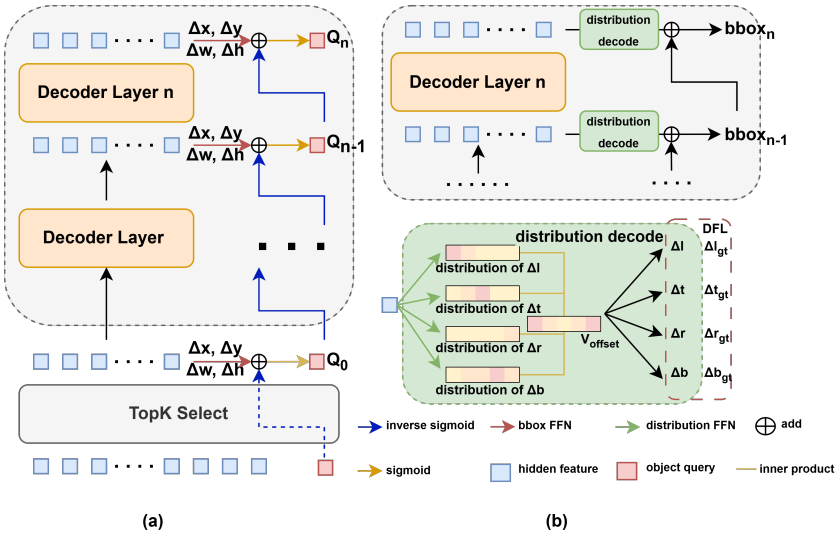

Figure 3. (**a**) DETR's decoding indirectly updates the prediction box by repeatedly updating it on the domain of the sigmoid function, which has inconsistency issues for multi-scale objects. (**b**) The distribution-based DETR decoder eliminates the need for repetitive sigmoid and inverse sigmoid calculations by employing probability distributions to represent the deviation between predicted and ground truth boxes, enhancing robustness in scenarios with dense occlusion.

3.2.2. Distribution-Based Decoder

The basic method is unstable because the value of Δx_{i+1} is highly dependent on Q_i. For instance, when dealing with small anchor boxes, their width and height approach zero, resulting in a larger update for Δx_{i+1} due to the maximal derivative of the Sigmoid function near zero. Conversely, updates for large anchor boxes are more stable. In conclusion, the basic method provides inconsistent updates for multi-scale Q_i, a common phenomenon in remote sensing. Addressing this inconsistency could enhance the utilization of end-to-end detection methods in remote sensing. Additionally, in densely overlapping scenes,

the ambiguity of anchor box boundaries leads to inaccurate predictions of Δx. In such cases, direct single-parameter regression by the model lacks robustness.

To alleviate this problem, we use discrete distributions to model a single parameter. By applying clamping, scaling, and sampling to the Sigmoid function, the continuous offset in the range $[-1, 1]$ is discretized into several offset categories using the following equation:

$$V_{\text{offset}} = s \cdot \left(\text{MinMaxScaler} \left(\log \frac{a}{1-a} \right) \right). \tag{4}$$

The potential offset V_{offset} is generated by uniformly sampling points a on the interval $[0.5, 1)$. As depicted in Figure 4, the sparsity of V_{offset} becomes more pronounced as the offset distance increases from 0. Here, s is the scaling factor, which varies with the feature map size and is usually set to 1. Drawing inspiration from FCOS [5], we utilize the anchor point and its distances to the four boundaries to describe the position of the box comprehensively. The probabilities associated with each offset are estimated by the feedforward network $\text{FFN}_{\text{distr}}$, and the final regression parameter values are obtained through integration. This process, referred to as distribution decoding and illustrated in Figure 3b, can be expressed as

$$\Delta x_{i+1} = \text{Softmax}(\text{FFN}_{\text{distri}}(F_{i+1}^{hide})) \cdot V_{\text{offset}}. \tag{5}$$

It is obtained by multiplying each offset by its predicted score and integrating. Consequently, the iterative update of the object query layer during the decoding phase is defined as

$$Q_{i+1} = \text{Clamp}(\Delta x_{i+1} + Q_i), \tag{6}$$

In this method, the decoder does not need to predict the update value on the Sigmoid domain, but directly updates the box. This approach not only reduces the reliance on sigmoid operations, but also enhances the model's ability to interpret occluded scenes. Although the discussion focuses on width and height offsets, the underlying concept is equally applicable to offsets in the left, right, up, and down directions relative to the anchor point.

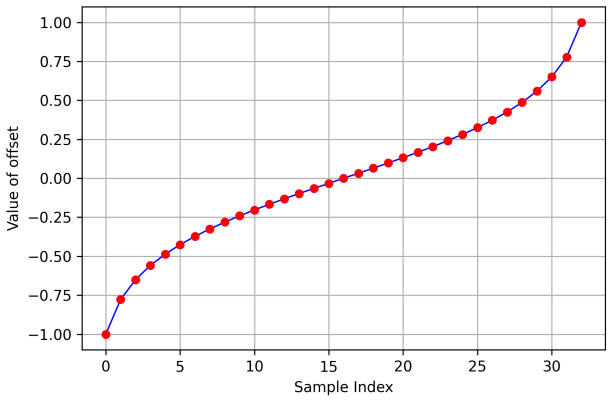

Figure 4. Schematic diagram of the numerical distribution of V_{offset}. Each red scatter point represents a potential offset, with the negative segment indicating the distance from the anchor point to the left and upper boundaries.

3.2.3. Distribution Focal Loss

As previously mentioned, predicting a single offset is not robust, leading us to draw inspiration from Generalized Focal Loss (GFL) [34]. Our aim is to align the discrete distribution of offsets closely with the Dirac distribution. Specifically, we seek to attain the highest possible score at the two pre-set offsets nearest to the target offset while ensuring

the integral sum approximates the target offset. Upon implementation, when the model predicts a boundary with high confidence, it will resemble the Dirac distribution. Conversely, when predicting an uncertain or indistinct boundary, the distribution will resemble a normal distribution because the model has learned that the offset should be near the most accurate location. The uncertainty at both extremes can be counterbalanced by positive and negative adjustments, thereby maintaining a relatively stable overall integral value.

In light of this, we introduce the Distribution Focal Loss (DFL), which is designed to facilitate the convergence of the distribution towards the Dirac distribution. We denote the true offset as Δx_{gt}, with its neighboring discrete values to the left and right represented by Δx_{gt}^{-} and Δx_{gt}^{+}, respectively. By computing the probabilities for predicting all possible discrete points of Δx, we employ cross-entropy loss to refine the adjustment values of the target boxes. The DFL can be formulated as follows:

$$DFL(\Delta x, \Delta x_{gt}) = (\Delta x_{gt} - \Delta x_{gt}^{-}) \times \mathrm{CE}(\Delta x, \Delta x_{gt}^{+}) + (\Delta x_{gt}^{+} - \Delta x_{gt}) \times \mathrm{CE}(\Delta x, \Delta x_{gt}^{-}), \quad (7)$$

where CE denotes the cross-entropy loss. The Distribution Focal Loss serves to supervise the offset between the decoding layers of the boxes, thereby effectively achieving the goal of residual learning in stacked decoding layers. Furthermore, this approach unifies offsets and object queries into a singular numerical space, in contrast to the basic method.

In Figure 5, we display various distribution scenarios that emerged during the training phase of the model. The probability density function in the first row exhibits a close alignment with the ground truth value, resembling a Dirac function in shape. In the second row, the model's localization appears less certain, resulting in object boundary offsets tending to avoid larger values. This behavior contributes to enhancing the robustness of the model, particularly when combined with the use of the Hungarian algorithm.

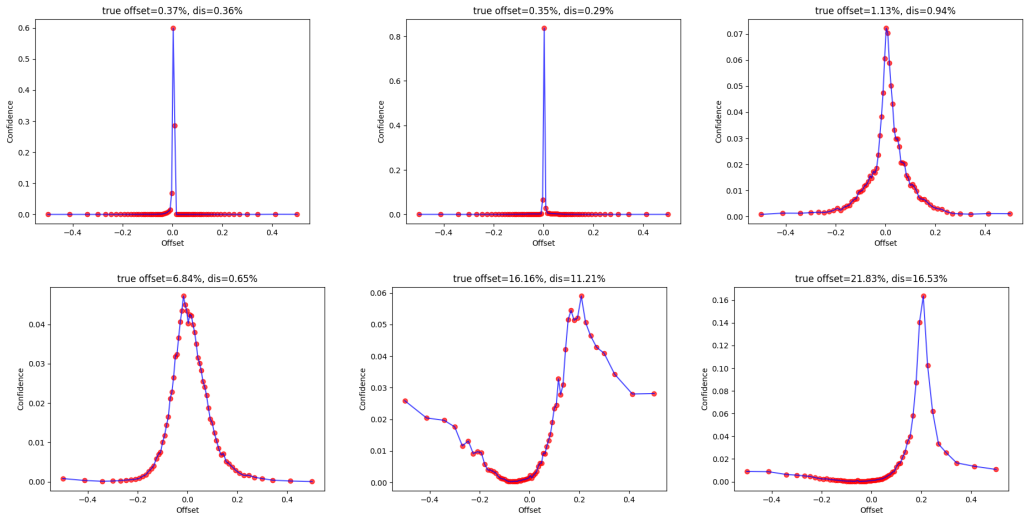

Figure 5. Schematic diagram of the offset distribution. Assuming that regression parameters adhere to the Dirac distribution, we represent the object boundary offset as the integral of the distribution to bolster the model's capability in localizing objects within occluded scenes.

3.3. High-Quality Query Selection Module

3.3.1. Basic Method

In DETR, object queries are conventionally represented as boxes within the image plane. These boxes are initialized by merging static anchor boxes with learned offsets. The initialization process is as follows: initially, a linear projection layer W_{score} is applied to

map the feature dimension to the number of categories present in the dataset, facilitating the identification of the maximum response value from the encoder's output feature f_{enc}.

$$f_{score} = \text{Max}(f_{enc} \cdot W_{score}). \tag{8}$$

Subsequently, the top-k encoder features with the highest response values are selected, where index denotes the indices of these top-k points within f_{enc}.

$$\text{Index} = topK(f_{score}). \tag{9}$$

We utilize the calculated index to retrieve the selected feature f'_{enc}, referred to as a token, and initialize the object query using this token,

$$Q_0 = \sigma(FFN_{enc}(f'_{enc}) + \widehat{\sigma}(A_{static})), \tag{10}$$

where FFN_{enc} denotes a feedforward network responsible for learning the offset between anchors and the ground truth, and A_{static} represents the statically initialized anchors based on the token's image position. Each A_{static} consists of four floating-point numbers ranging from 0.0 to 1.0, indicating the anchor box's central position, width, and height. The functions σ and $\widehat{\sigma}$ represent the sigmoid and inverse sigmoid operations, respectively. The resulting initialized object query is expressed as Q_0. The fundamental approach prioritizes the response scores of local features across all categories, focusing on the selection of foreground features. However, in remote sensing applications, numerous objects are often present within a single image. Without increasing the number of object queries, features corresponding to small-scale or occluded objects may be underrepresented, leading to missed detections. Furthermore, even with an adequate number of object queries, inaccuracies may arise during the subsequent matching process performed by the Hungarian algorithm. DETR employs the Hungarian matching algorithm to establish an end-to-end one-to-one correspondence between prediction and label boxes. It constructs a cost matrix for bipartite matching by calculating the matching cost for all pairs of prediction and label boxes. Let $y = \{(c_j, b_j)\}_{j=1}^{N}$ denote the set of true label boxes, where c_j and b_j represent the category and coordinates of label box j, respectively. Similarly, $\hat{y} = \{(\hat{c}_i, \hat{b}_i)\}_{i=1}^{n^{obj}}$ is the set of prediction boxes, with $\hat{c}_i$ and $\hat{b}_i$ representing the category and coordinates of prediction box i. The cost for matching prediction box i with label box j is given by

$$\text{Cost}(i, j) = \mathcal{L}_{\text{cls}}(\hat{c}_i, c_j) + \lambda_{\text{box}}\mathcal{L}_{\text{box}}(\hat{b}_i, b_j) + \lambda_{\text{iou}}\mathcal{L}_{\text{iou}}(\hat{b}_i, b_j),$$

where λ_{box} and λ_{iou} are weight hyperparameters for the bounding box and IoU losses, typically set to $\lambda_{\text{box}} = 5$ and $\lambda_{\text{iou}} = 2$. To form a square cost matrix, $n^{obj} - N$ columns of zero vectors are added, representing pseudo-labels that do not affect the final matching result. The matching process, as shown in Figure 6, involves assigning one-to-one correspondences to positive samples (e.g., rows marked in blue).

As depicted in Figure 6, the cost matrix is populated by calculating the discrepancies between predicted values and actual ground truth. Each ground truth is optimally matched with a prediction to minimize the total cost, which includes both classification and regression losses. The cost function's structure, which combines classification and regression losses, can lead to an imbalance between these processes. Object queries linked to objects with indistinct features are often discarded due to classification challenges, potentially resulting in incorrect matching with larger, neighboring objects.

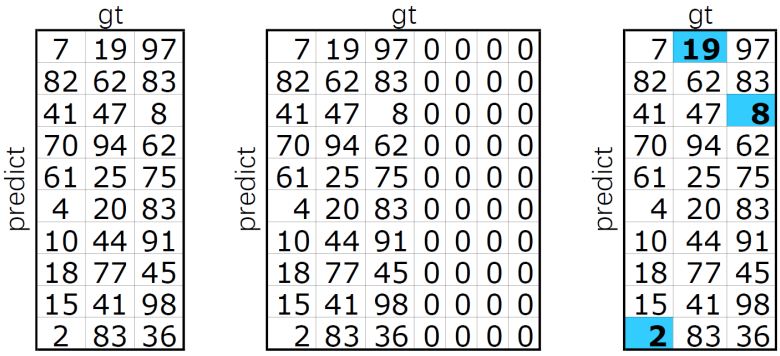

Figure 6. Cost matrix schematic diagram. This figure illustrates the calculation and population of all possible losses between predictions and ground truth into a square matrix, with the Hungarian algorithm identifying the minimal one-to-one matching solution. Note: the values in the figure are randomly generated and are for illustrative purposes only, not representative of experimental data.

3.3.2. HQQS Module

In remote sensing detection of dense targets, noise in the image can lead to low response scores, potentially causing the elimination of valid object queries in the initial phase. During subsequent decoding, these queries may be incorrectly matched with nearby simple targets, causing further learned and updated offsets. This mismatch primarily arises because the conventional method initializes object queries based on response scores, which produce numerous redundant object queries for large objects with prominent features. When objects are densely packed and anchor boxes overlap, there is a risk of erroneous matching with smaller targets. In conclusion, the conventional method exhibits limited interaction between classification and regression tasks.

To address the identified inconsistency in the query selection process, we have developed a High-Quality Query Selection (HQQS) module, which is designed to improve the overall effectiveness of query selection. Our module employs a feedforward network to process all input tokens, x_{all}, as represented by the following equation:

$$X_{box} = \sigma(FFN_{enc}(x_{all}) + \widehat{\sigma}(anchor_{static})). \tag{11}$$

Following this, we utilize Non-Maximum Suppression (NMS) to simultaneously consider the uniqueness of predicted positions and their associated confidence levels,

$$Q_0 = TopK(QS(X_{box}, X_{score})). \tag{12}$$

The final step involves selecting the top K tokens from the retained boxes to initialize the query. It is important to note that within the HQQS module, QS is implemented with a class-agnostic approach, using an Intersection over Union (IoU) filtering threshold of 0.8. This module operates independently of the post-processing NMS used in one-stage detectors, adhering to a fixed computational flow aimed at eliminating redundant high-confidence features. This strategy enhances query diversity and improves the detection accuracy for objects that are partially obscured or of small scale. We have deliberately chosen lenient criteria for suppression to avoid incorrect suppression in challenging scenarios, while simultaneously ensuring the computational efficiency of the module.

3.4. Short-Circuit Training Decoder

The lack of accurate initial positions in object queries presents a significant challenge for optimizing the decoding layer and exacerbates the inherent instability of using the Hungarian algorithm for positive sample assignment. To address this, we have fil-

tered object queries to remove redundancy, ensuring that challenging objects are also correctly initialized.

The Hungarian algorithm employed by DETR is limited in generality and robustness, as it relies solely on a manually designed weighted sum of the prediction and label losses for the classification and regression tasks. Additionally, DETR guarantees an equal number of positive samples and labels in each batch of training data, whereas staged detection algorithms typically have hundreds of positive samples per batch. This discrepancy hinders the convergence and learning process of the decoder. Beyond DETR, developing an accurate matching mechanism to support the Hungarian algorithm can improve the stability and convergence speed of the original method.

Drawing inspiration from Task-aligned One-stage Object Detection [24], we have incorporated additional auxiliary detection heads and innovative assignment strategies into our model. These enhancements facilitate a one-to-many matching between the encoder's predictions and ground truth labels. Specifically, the quality of the predicted bounding boxes is assessed based on the confidence of the correctly predicted class as well as the IoU metric, thereby yielding more precise soft labels. For instance, a prediction with inaccurate localization will have its classification label correspondingly adjusted downwards. A comprehensive description of the assignment process is presented in Algorithm 1. Line 3 of the algorithm ensures that the object queries remain within the bounds of the annotated box. Lines 4 and 5 evaluate the accuracy of the object query based on Intersection over Union (IoU), selecting the top k entries. Lines 6 and 7 calculate the prediction quality for two tasks, IoU and classification confidence, and integrate these to formulate the *Align metric*. Lines 8 and 9 apply the *Align metric* to the original one-hot label to establish an accurate positive and negative sample matching mechanism for subsequent training.

Algorithm 1 Task-aligned assignment algorithm

Input:
- G_b: a set of ground truth boxes
- G_c: a set of class labels
- P_b: a set of predicted boxes
- P_c: a set of predicted class scores
- A_p: a set of static anchor points
- k: a hyperparameter with a default value of 10
- α: a hyperparameter with a default value of 1
- β: a hyperparameter with a default value of 6

Output:
- T_b: a set of target boxes
- T_c: a set of target scores

1: Initialize the output sets $T_b \leftarrow \emptyset$ and $T_c \leftarrow \emptyset$.
2: **for** each ground truth $g_b \in G_b$ and corresponding $g_c \in G_c$ **do**
3: $c_b \leftarrow$ Select candidates from P_b where A_p is within g_b based on L2 distance.
4: Compute IoU between c_b and g_b: $D_g = IoU(c_b, g_b)$.
5: $c^k \leftarrow$ Select the top k candidates for g_b from c_b according to D_g.
6: Let c_c^k and c_d^k denote the predicted scores and IoU between c^k and g_b, respectively.
7: *Align metric* $\leftarrow c_c^{k^\alpha} \times c_d^{k^\beta}$;
8: $C_{1-hot} \leftarrow$ get_one_hot_labels($\mathbf{c}^k, g_c$);
9: $C_c \leftarrow C_{1-hot} \times$ *Align metric*;
10: $T_b \leftarrow T_b \cup C_b$;
11: $T_c \leftarrow T_c \cup C_c$;
12: **end for**
13: **return** T_b, T_c.

Subsequent to this step, the need for Hungarian matching is obviated, and loss calculation is performed on this new branch,

$$\mathcal{L} = \mathcal{L}_{ori} + \lambda \mathcal{L}_{aux}, \tag{13}$$

where the loss is separated into two components: the original Hungarian matching loss and the auxiliary loss. The scalar λ represents a balance parameter, and our findings indicate that setting it to 1 yields favorable results. Furthermore, the loss can be decomposed into smooth L1 loss [37], Generalized Intersection over Union (GIoU) loss [38,39], and cross-entropy loss, with the following weighting scheme:

$$\mathcal{L} = 5\mathcal{L}_{box} + 2\mathcal{L}_{iou} + \mathcal{L}_{cls}. \tag{14}$$

Table 1. Comparison of model performance on the COCO validation set at input size 800 × 1333 pixels. DHQ-DETR achieves a 0.6% increase in AP compared to baseline.

Model	Epochs	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L	Params	GFLOPS
DETR-R50 [16]	500	42.0	62.4	44.2	20.5	45.8	61.1	41M	86
Anchor DETR-R50 [40]	50	42.1	63.1	44.9	22.3	46.2	60.0	39M	–
Conditional DETR-R50 [41]	50	40.9	61.8	43.3	20.8	44.6	59.2	44M	90
DAB-DETR-R50 [18]	50	42.2	63.1	44.7	21.5	45.7	60.3	44M	94
DN-DETR-R50 [19]	50	44.1	64.4	46.7	22.9	48.0	63.4	44M	94
Align-DETR-R50 [42]	50	46.0	64.9	49.5	25.2	50.5	64.7	42M	94
RT-DETR-R50 [22]	72	53.1	71.3	57.7	34.8	58.0	70.0	42M	136
YOLOv5 L [10]	300	49.0	67.3	–	–	–	–	46M	109
YOLOv7 L [11]	300	51.2	69.7	55.5	35.2	55.9	66.7	36M	104
YOLOv8 L [12]	300	52.9	69.8	57.5	**35.3**	58.3	69.8	43M	165
DETR-R101 [16]	500	43.5	63.8	46.4	21.9	48.0	61.8	60M	152
Anchor DETR-R101 [40]	50	43.5	64.3	46.6	23.2	47.7	61.4	58M	–
Conditional DETR-R101 [41]	50	42.8	63.7	46.0	21.7	46.6	60.9	63M	156
DAB-DETR-R101 [18]	50	43.5	63.9	46.6	23.6	47.3	61.5	63M	174
DN-DETR-R101 [19]	50	45.2	65.5	48.3	24.1	49.1	65.1	63M	174
Align-DETR-R101 [42]	50	46.9	65.5	50.9	25.6	51.9	66.0	61M	174
DETR-DC5-R50 [16]	500	43.3	63.1	45.9	22.5	47.3	61.1	41M	187
Anchor DETR-DC5-R50 [40]	50	44.2	64.7	47.5	24.7	48.2	60.6	39M	151
Conditional DETR-DC5-R50 [41]	50	43.8	64.4	46.7	24.0	47.6	60.7	44M	195
DAB-DETR-DC5-R50 [18]	50	44.5	65.1	47.7	25.3	48.2	62.3	44M	202
DN-DETR-DC5-R50 [19]	50	46.3	66.4	49.7	26.7	50.0	64.3	44M	202
Align-DETR-DC5-R50 [42]	50	48.3	66.7	52.5	29.7	52.8	65.9	42M	200
DETR-DC5-R101 [16]	500	44.9	64.7	47.7	23.7	49.5	62.3	60M	253
Anchor DETR-DC5-R101 [40]	50	45.1	65.7	48.8	25.8	49.4	61.6	58M	–
Conditional DETR-DC5-R101 [41]	50	45.0	65.5	48.4	26.1	48.9	62.8	63M	262
DAB-DETR-DC5-R101 [18]	50	45.8	65.9	49.3	27.0	49.8	63.8	63M	282
DN-DETR-DC5-R101 [19]	50	47.3	67.5	50.8	28.6	51.5	65.0	63M	282
Align-DETR-DC5-R101 [42]	50	49.3	67.4	53.7	30.6	54.3	66.4	61M	280
DHQ-DETR	72	**53.7**	**71.6**	**57.9**	34.7	**58.4**	**70.6**	43M	154

The bold data represent the best results of different evaluation indicators.

4. Results

Dataset and evaluation metrics. Our experimental analysis was conducted using three benchmark datasets: DOTAv1.0, VisDrone, and MS COCO 2017 [36,43,44]. These datasets span both natural image (COCO) and remote sensing image domains (DOTA and VisDrone), offering a comprehensive evaluation of our method's versatility. The MS COCO 2017 dataset includes 115,000 training images and 5000 validation images, featuring 80 object categories commonly seen in daily life. Objects in COCO are typically larger relative to the image resolution, with more structured backgrounds and less scale variation

compared to remote sensing datasets. In contrast, DOTAv1.0 and VisDrone represent challenging remote sensing tasks. DOTAv1.0 focuses on aerial imagery, characterized by large-scale scenes with diverse object scales, dense distributions, and complex, unstructured backgrounds. VisDrone, captured from drone perspectives, emphasizes dense and overlapping small objects, such as vehicles and pedestrians, within cluttered urban environments. Both datasets highlight the difficulties of small object detection, high-density scenarios, and intricate backgrounds, which are less prominent in COCO. We evaluate performance using the standard Average Precision (AP) metric, with AP_{50} denoting AP at an IoU threshold of 50%, and AP representing the average AP across IoU thresholds ranging from 0.5 to 0.95. By leveraging these datasets, we validate our method's ability to adapt to both natural and remote sensing domains, demonstrating robust performance across diverse detection challenges.

Table 2. Results on the DOTAv1.0 test set. Our model achieves a 1.3 increase in AP over the baseline. ✓indicates that the corresponding method or data was used, while × indicates that it was not used.

Model	Extra Data	AP	AP_{50}	AP_{75}
YOLOv5 (2020) [10]	×	49.0	73.0	50.9
YOLOv8 (2023) [12]	×	52.9	74.5	56.1
DETR (2020) [16]	×	46.7	72.3	49.5
DN-DETR (2022) [19]	×	53.1	78.2	57.5
RT-DETR (2023) [22]	×	53.0	79.0	57.8
DecoupleNet D2 (2024) [45]	✓	-	78.0	-
PP-YOLOE-R-l (2022) [46]	✓	-	80.0	-
MAE + MTP (2024) [47]	✓	-	80.7	-
LSKNet (2024) [48]	✓	-	81.6	-
Strip R-CNN (2025) [49]	✓	-	**82.3**	-
DHQ-DETR	×	**54.3**	81.5	**58.9**

The bold data represent the best results of different evaluation indicators.

Implementation details. We adopt RT-DETR as our baseline model. During training, the loss computed in the short-circuit training decoder is exclusively used to update the decoder's parameters, with the gradients for the backbone and encoder being discarded. The training involves the use of four NVIDIA RTX A6000 GPUs for parallel processing. The AdamW optimizer is utilized with a learning rate of 1×10^{-3}, a momentum of 0.937, and a weight decay of 5×10^{-4}. The training process lasts for 72 epochs on the COCO dataset, accompanied by a linear learning rate decay that reduces it to 10% of the initial value. Unless otherwise stated, the decoder stage is initialized with 300 queries, and the model's training parameters remain unchanged throughout the experiments.

4.1. Main Results

In the present section, we undertake an empirical analysis to assess the efficacy and generalization ability of the DHQ-DETR model. It should be noted that all models involved in this research process input images by resizing them to a consistent dimension of 800 × 1333 pixels. We present a comparative evaluation of several state-of-the-art methods on the frequently utilized COCO dataset for object detection in Table 1. Our empirical results indicate a significant improvement of 0.6 in AP on the COCO validation set. This advancement can be ascribed to the optimized offset distribution, which enhances the quality of representation, refines the precision of object localization, and consequently achieves superior performance metrics.

DOTAv1.0 is a dataset designed for remote sensing target detection, characterized by challenges such as high target density and significant scale variations. In Table 2, comparative analysis on the benchmark test set revealed that the enhanced method demonstrated greater robustness and higher detection accuracy under these complex conditions. Furthermore, when compared with the most advanced detection methods, the proposed method achieved performance levels nearly equivalent to state-of-the-art techniques without utilizing additional data. The Visdrone dataset, which consists of drone aerial imagery featuring dense objects, poses a substantial challenge for detectors in terms of dense multi-scale detection. For this investigation, we evaluated detection algorithms specifically enhanced for multi-scale detection, resulting in a notable 1.4-point increase in AP on the test set (Table 3). Our experimental data were derived from the competitive results presented by Zhu et al. [36]. It is evident from our findings that the DETR does not exhibit a clear strength in detecting small objects, whereas our enhanced method significantly outperforms it in this regard.

Table 3. Results on the Visdrone test set. Our model achieves a 1.4-point increase in AP over the baseline and demonstrates enhanced performance compared to models optimized for the object scale characteristics of Visdrone.

Model	AP	AP_{50}	AP_{75}
Cascade R-CNN [1]	16.0	31.9	15.0
HTC-drone [50]	22.6	45.2	20.0
Libra-HBR [51]	25.6	48.3	24.0
HRDet+ [52]	28.4	54.5	26.1
S + D [1,53]	28.6	51.0	28.3
ACM-OD [3,54]	29.1	54.1	27.4
DPNet [1,55]	29.6	54.0	28.7
RRNet [56]	29.1	55.8	27.2
RetinaNet [7]	11.8	21.3	11.6
CornerNet [57]	17.4	34.1	15.8
YOLOv3 [58]	17.8	37.3	15.0
TridentNet [59]	22.5	43.3	20.5
CNAnet [60]	26.4	48.0	25.5
EHR-RetinaNet [7]	26.5	48.3	25.4
CN-DhVaSa [61]	27.8	50.7	26.8
DETR (2020) [16]	23.1	39.8	25.7
DN-DETR (2022) [19]	31.4	51.6	26.8
RT-DETR (2023) [22]	31.0	50.2	26.9
CZ Det (2023) [62]	32.2	54.9	**31.2**
DHQ-DETR	**32.4**	**55.4**	30.0

The bold data represent the best results of different evaluation indicators.

4.2. Ablation Studies

4.2.1. Distribution-Based Location Offset

In the ablation study on the distribution-based location offset, we explored various sampling levels of distribution discretization. The results indicate an improvement of 0.7% in the AP when the distribution representation is combined with DFL. The details are presented in Table 4.

To rigorously assess the impact of the distribution-based location offset on the model's positioning accuracy, an ablation study was meticulously designed. We employed non-uniform sampling within the range of $[-0.5, 0.5]$ for the offset, with the "Sampling level" denoting the number of sampling points. The model was trained uniformly for 36 epochs.

As indicated in Table 4, the model incorporating the distribution-based location offset retained satisfactory detection efficacy even in the absence of the Distance-Friendly Loss (DFL). However, the addition of DFL enhanced the AP by 0.7%. This enhancement can be attributed to the fact that DFL can promote explicit residual learning between decoding layers, which makes the predicted discrete distribution converge to the Dirac distribution. Therefore, the model has stronger explanatory power when dealing with fuzzy boundaries.

Table 4. Results of ablation study on distribution-based location offset. ✓indicates that the corresponding method or data was used, while × indicates that it was not used.

Model	Sampling Level	DFL	Epochs	AP	AP_{50}
RT-DETR [22]	×	×	36	48.7	67.1
DHQ-DETR	16	×	36	48.5	66.6
DHQ-DETR	32	×	36	48.8	67.0
DHQ-DETR	64	×	36	48.7	67.1
DHQ-DETR	16	✓	36	49.1	67.3
DHQ-DETR	32	✓	36	**49.4**	**67.5**
DHQ-DETR	64	✓	36	49.3	67.3

The bold data represent the best results of different evaluation indicators.

4.2.2. HQQS Module

An in-depth ablation study was conducted on the HQQS module, and the quantitative results are presented in Table 5. We compared three methods: (1) Vanilla, which utilizes standard cross-entropy as the classification loss; (2) IoU-aware, which integrates IoU into the classification loss to enhance the selection of queries; and (3) HQQS module, which considers IoU during the query selection phase to initialize features with spatial correspondence. We evaluated two metrics on the COCO validation set: the proportion of encoder feature classification scores exceeding 0.5 ($Prop_{cls}$), and the mean maximum IoU between encoder-predicted boxes and ground-truth instances ($MeanIoU$). The findings demonstrate that the HQQS module not only matches the $Prop_{cls}$ performance of the IoU-aware method, but also significantly improves the spatial correspondence of initialized queries. The incorporation of the HQQS module led to a 0.8% increase in AP.

Table 5. Results of the ablation study on the HQQS module. The symbol $Prop_{cls}$ denotes the proportion of selected query feature scores greater than 0.5, while $MeanIoU$ represents the average IoU with instances that have the highest IoU. The HQQS module improves the precision and comprehensiveness of object-based queries by 10%. ✓indicates that the corresponding method or data was used, while × indicates that it was not used.

Model	IoU-Aware [63]	HQQS	AP	$Prop_{cls}$	$Mean_{IoU}$
RT-DETR [22]	×	×	47.9	0.35	0.47
RT-DETR [22]	✓	×	48.7	**0.82**	0.45
RT-DETR [22]	✓	✓	**49.5**	0.79	**0.58**

The bold data represent the best results of different evaluation indicators.

4.3. Assignment Strategies

In the decoder, the employment of an auxiliary detection head alongside a judicious assignment strategy facilitates a one-to-many mapping between object queries and ground truths. Nonetheless, there is considerable variation in the assignment strategies employed by different models. For example, RetinaNet relies on anchors and anchor-based offsets, Faster R-CNN uses region proposal networks, ATSS incorporates statistical measures such as mean and variance to distinguish positive and negative samples, and FCOS focuses

on the central position of the bounding box. Our approach utilizes the same detection head as YOLOv8. To gauge the effectiveness of our proposed assignment strategy, we conducted a comparative experiment. Table 6 illustrates that the adoption of our additional assignment strategies, as opposed to the Vanilla approach, can bolster the positive supervision of the decoder, thereby enhancing performance. Notably, our strategy yielded the most substantial improvement in AP, with a 2.1% increase, by effectively aligning classification and regression tasks and providing more precise segmentation of positive and negative samples.

Table 6. Results of the ablation study on assignment strategies. Our detection head and assignment strategy demonstrated the most favorable outcomes, improving the AP by 2.1% compared to the baseline.

Assignment Strategies	Epochs	AP	AP_{50}
Vanilla	36	48.7	67.1
RetinaNet [7]	36	49.6	67.9
Faster R-CNN [3]	36	49.9	68.9
FCOS [5]	36	50.1	68.6
ATSS [64]	36	50.4	68.9
Ours	36	**50.8**	**69.0**

The bold data represent the best results of different evaluation indicators.

5. Discussion

The DHQ-DETR model effectively tackles the intricate challenge of detecting dense multi-scale objects in remote sensing imagery, demonstrating impressive performance across specific datasets. However, this approach has several limitations that require further investigation. Addressing these limitations could pave the way for future research aimed at improving the model's robustness and applicability.

5.1. Limitations

Despite advancements in the convergence speed of DHQ-DETR and enhanced detection performance of DETR for densely populated small targets, which expand the applicability of end-to-end detection methods in remote sensing, several limitations persist. A major challenge in applying natural image object detection techniques to remote sensing imagery is the unique characteristics of these datasets, including resolution constraints and complex environmental factors. Specifically, scenes with vast areas and small objects pose significant difficulties for natural object detection models, which often struggle to capture fine-grained details. This limitation frequently leads to missed detections or false positives for smaller targets. To address these challenges, divide-and-conquer strategies are commonly employed. For instance, when processing datasets like DOTAv1.0, large 8k × 8k images are typically cropped into smaller sections. This approach helps fit the data within memory constraints while preserving image details. However, although these strategies alleviate computational limitations, they inadvertently disrupt the global spatial context—a crucial factor for accurately identifying small or densely packed objects in remote sensing applications. Moreover, even though DHQ-DETR demonstrates outstanding performance on datasets such as COCO val2017, DOTAv1.0, and Visdrone, its generalizability to untested datasets or real-world conditions remains unaddressed. Additional challenges include managing occlusions, variable illumination, and diverse object appearances, which are particularly significant in remote sensing scenarios. Extreme environmental conditions, such as overlapping objects and inconsistent lighting, further impede detection performance, highlighting the need for robust and adaptive solutions tailored to the unique requirements of remote sensing imagery.

5.2. Future Research Directions

To address these limitations, future research should focus on several strategies. Optimization through model simplification or pruning could reduce computational demands without significant performance trade-offs, facilitating deployment in resource-constrained environments. Techniques such as transfer learning and domain adaptation could enhance the model's robustness and adaptability across diverse and unseen datasets. To better handle highly occluded and variable scenes, further architectural modifications and data augmentation techniques should be explored. Scalability issues might be mitigated by investigating hierarchical or multi-resolution processing methods, thereby enhancing the model's capability to handle large, high-resolution images typical in remote sensing. Finally, integrating complementary data sources, such as LiDAR or hyperspectral imaging, could provide additional contextual information, potentially improving detection accuracy in challenging scenarios. In conclusion, while the DHQ-DETR model represents a significant advancement in dense multi-scale object detection, targeted research addressing its limitations could further enhance its utility and application, leading to broader adoption across a range of remote sensing tasks.

6. Conclusions

This study attempts to address the complex challenges of dense multi-scale object detection in aerial images using a distribution-based box offset modeling approach in the remote sensing field. We introduce the Distribution Focus Loss (DFL) to facilitate residual learning between decoded outputs and ground truth labels, thereby enhancing the model's ability to accurately localize objects in densely occluded scenes. Additionally, to initialize object queries with precise spatial relationships, we present a High-Quality Query Selection (HQQS) module. To accelerate convergence and improve decoder performance, we employ an auxiliary head with an innovative assignment strategy that enables one-to-many matching during training, providing additional positive supervision. Our experimental results confirm the effectiveness of the proposed DHQ-DETR model, achieving an AP of 53.7% on the COCO val2017 dataset, 54.3% on the DOTAv1.0 test set, and 32.4% on the Visdrone test set, surpassing other existing detectors of similar scale.

Author Contributions: Conceptualization, C.L.; Data curation, J.Z. and Y.X.; Funding acquisition, J.Z.; Methodology, C.L. and Y.X.; Project administration, J.Z.; Supervision, J.Z.; Validation, B.H. and Y.X.; Visualization, C.L.; Writing—original draft, C.L. and B.H.; Writing—review and editing, J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under Grants 62076137.

Data Availability Statement: All data included in this study are available upon request by contact with the corresponding author.

Acknowledgments: This work received partial funding from the National Natural Science Foundation of China under Grants 62076137 (Corresponding author: Jianwei Zhang).

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

Abbreviations

The following abbreviations are used in this manuscript:

DETR	Detection transformer
NMS	Non-Maximum Suppression
DHQ	Distributed and high-quality object query

HQQS	High-quality query selection module
FFN	Feedforward network
DFL	Distribution focus loss

References

1. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6154–6162.
2. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; Number 1, pp. 1440–1448.
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 1137–1149. [CrossRef] [PubMed]
4. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
5. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
6. Long, X.; Deng, K.; Wang, G.; Zhang, Y.; Dang, Q.; Gao, Y.; Shen, H.; Ren, J.; Han, S.; Ding, E.; et al. PP-YOLO: An Effective and Efficient Implementation of Object Detector. *arXiv* **2020**, arXiv:2007.12099.
7. 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; pp. 2980–2988.
8. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.E.; Fu, C.; Berg, A.C. SSD: Single Shot MultiBox Detector. *arXiv* **2015**, arXiv:1512.02325. [CrossRef]
9. Bochkovskiy, A.; Wang, C.Y.; Liao, H.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
10. Jocher, G. YOLOv5 Release v7.0. 2022. Available online: https://github.com/ultralytics/yolov5/tree/v7.0 (accessed on 1 March 2023).
11. Wang, C.Y.; Bochkovskiy, A.; Liao, H.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-time Object Detectors. *arXiv* **2022**, arXiv:2207.02696.
12. Jocher, G. YOLOv8. 2023. Available online: https://github.com/ultralytics/ultralytics/tree/main (accessed on 1 March 2023).
13. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Improving Object Detection With One Line of Code. *arXiv* **2017**, arXiv:1704.04503. [CrossRef]
14. Zhou, P.; Zhou, C.; Peng, P.; Du, J.; Sun, X.; Guo, X.; Huang, F. NOH-NMS: Improving Pedestrian Detection by Nearby Objects Hallucination. *arXiv* **2020**, arXiv:2007.13376. [CrossRef]
15. Liu, S.; Huang, D.; Wang, Y. Adaptive NMS: Refining Pedestrian Detection in a Crowd. *arXiv* **2019**, arXiv:1904.03629. [CrossRef]
16. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 1, 2, 4, 6.
17. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv* **2021**, arXiv:2010.04159. [CrossRef]
18. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. *arXiv* **2022**, arXiv:2201.12329.
19. Li, F.; Zhang, H.; Liu, S.; Guo, J.; Ni, L.M.; Zhang, L. DN-DETR: Accelerate DETR training by introducing query denoising. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 2, 3, 5, 6.
20. Chen, Q.; Chen, X.; Wang, J.; Zhang, S.; Yao, K.; Feng, H.; Han, J.; Ding, E.; Zeng, G.; Wang, J. Group DETR: Fast DETR Training with Group-Wise One-to-Many Assignment. *arXiv* **2023**, arXiv:2207.13085. [CrossRef]
21. Roh, B.; Shin, J.; Shin, W.; Kim, S. Sparse DETR: Efficient End-to-End Object Detection with Learnable Sparsity. *arXiv* **2021**, arXiv:2111.14330. [CrossRef]
22. Lv, W.; Zhao, Y.; Xu, S.; Wei, J.; Wang, G.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. DETRs Beat YOLOs on Real-time Object Detection. *arXiv* **2023**, arXiv:2304.08069. [CrossRef]
23. Zong, Z.; Song, G.; Liu, Y. DETRs with Collaborative Hybrid Assignments Training. *arXiv* **2023**, arXiv:2211.12860. [CrossRef]
24. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. TOOD: Task-aligned One-stage Object Detection. *arXiv* **2021**, arXiv:2108.07755. [CrossRef]
25. Kirillov, A.; Girshick, R.; He, K.; Dollár, P. Panoptic feature pyramid networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6399–6408.
26. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.

28. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
29. Hosang, J.H.; Benenson, R.; Schiele, B. A convnet for non-maximum suppression. *arXiv* **2015**, arXiv:1511.06437. [CrossRef]
30. Hosang, J.H.; Benenson, R.; Schiele, B. Learning non-maximum suppression. *arXiv* **2017**, arXiv:1705.02950. [CrossRef]
31. Solovyev, R.A.; Wang, W. Weighted Boxes Fusion: Ensembling boxes for object detection models. *arXiv* **2019**, arXiv:1910.13302. [CrossRef]
32. Choi, J.; Chun, D.; Kim, H.; Lee, H. Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving. *arXiv* **2019**, arXiv:1904.04620. [CrossRef]
33. He, Y.; Zhang, X.; Savvides, M.; Kitani, K. Softer-NMS: Rethinking Bounding Box Regression for Accurate Object Detection. *arXiv* **2018**, arXiv:1809.08545. [CrossRef]
34. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. *arXiv* **2020**, arXiv:2006.04388. [CrossRef]
35. Zheng, D.; Dong, W.; Hu, H.; Chen, X.; Wang, Y. Less is More: Focus Attention for Efficient DETR. *arXiv* **2023**, arXiv:2307.12612. [CrossRef]
36. Zhu, P.; Wen, L.; Du, D.; Bian, X.; Fan, H.; Hu, Q.; Ling, H. Detection and tracking meet drones challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 7380–7399. [CrossRef]
37. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv* **2013**, arXiv:1311.2524. [CrossRef]
38. Tychsen-Smith, L.; Petersson, L. Improving Object Localization with Fitness NMS and Bounded IoU Loss. *arXiv* **2017**, arXiv:1711.00164. [CrossRef]
39. Rezatofighi, S.H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.D.; Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. *arXiv* **2019**, arXiv:1902.09630. [CrossRef]
40. Wang, Y.; Zhang, X.; Yang, T.; Sun, J. Anchor detr: Query design for transformer-based detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 28 February–1 March 2022; pp. 2, 3, 6.
41. Meng, D.; Chen, X.; Fan, Z.; Zeng, G.; Li, H.; Yuan, Y.; Sun, L.; Wang, J. Conditional detr for fast training convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 10–17 October 2021; pp. 2, 3, 6.
42. Cai, Z.; Liu, S.; Wang, G.; Ge, Z.; Zhang, X.; Huang, D. Align-DETR: Improving DETR with Simple IoU-aware BCE loss. *arXiv* **2023**, arXiv:2304.07527. [CrossRef]
43. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A large-scale dataset for object detection in aerial images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3974–3983.
44. Lin, T.; Maire, M.; Belongie, S.J.; Bourdev, L.D.; Girshick, R.B.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. *arXiv* **2014**, arXiv:1405.0312. [CrossRef]
45. Lu, W.; Chen, S.B.; Shu, Q.L.; Tang, J.; Luo, B. DecoupleNet: A Lightweight Backbone Network With Efficient Feature Decoupling for Remote Sensing Visual Tasks. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4414613. [CrossRef]
46. Wang, X.; Wang, G.; Dang, Q.; Liu, Y.; Hu, X.; Yu, D. PP-YOLOE-R: An Efficient Anchor-Free Rotated Object Detector. *arXiv* **2022**, arXiv:2211.02386. [CrossRef]
47. Wang, D.; Zhang, J.; Xu, M.; Liu, L.; Wang, D.; Gao, E.; Han, C.; Guo, H.; Du, B.; Tao, D.; et al. MTP: Advancing Remote Sensing Foundation Model via Multitask Pretraining. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11632–11654. [CrossRef]
48. Li, Y.; Li, X.; Dai, Y.; Hou, Q.; Liu, L.; Liu, Y.; Cheng, M.M.; Yang, J. LSKNet: A Foundation Lightweight Backbone for Remote Sensing. *arXiv* **2024**, arXiv:2403.11735. [CrossRef]
49. Yuan, X.; Zheng, Z.; Li, Y.; Liu, X.; Liu, L.; Li, X.; Hou, Q.; Cheng, M.M. Strip R-CNN: Large Strip Convolution for Remote Sensing Object Detection. *arXiv* **2025**, arXiv:2501.03775. [CrossRef]
50. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid Task Cascade for Instance Segmentation. *arXiv* **2019**, arXiv:1901.07518. [CrossRef]
51. Jiangmiao, P.; Kai, C.; Jianping, S.; Huajun, F.; Wanli, O.; Dahua, L. Libra R-CNN: Towards Balanced Learning for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019.
52. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 3349–3364. [CrossRef]
53. Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *arXiv* **2016**, arXiv:1606.00915. [CrossRef] [PubMed]

54. Hong, S.; Kang, S.; Cho, D. Patch-Level Augmentation for Object Detection in Aerial Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 127–134. [CrossRef]
55. 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.
56. Chen, C.; Zhang, Y.; Lv, Q.; Wei, S.; Wang, X.; Sun, X.; Dong, J. RRNet: A Hybrid Detector for Object Detection in Drone-Captured Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 100–108. [CrossRef]
57. Law, H.; Deng, J. CornerNet: Detecting Objects as Paired Keypoints. *arXiv* **2018**, arXiv:1808.01244. [CrossRef]
58. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
59. Li, Y.; Chen, Y.; Wang, N.; Zhang, Z. Scale-Aware Trident Networks for Object Detection. *arXiv* **2019**, arXiv:1901.01892. [CrossRef]
60. Yang, B.; Xu, W.; Bi, F.; Zhang, Y.; Kang, L.; Yi, L. Multi-scale neighborhood query graph convolutional network for multi-defect location in CFRP laminates. *Comput. Ind.* **2023**, *153*, 104015. [CrossRef]
61. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as Points. *arXiv* **2019**, arXiv:1904.07850.
62. Meethal, A.; Granger, E.; Pedersoli, M. Cascaded Zoom-in Detector for High Resolution Aerial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023.
63. Zhang, H.; Wang, Y.; Dayoub, F.; Sünderhauf, N. VarifocalNet: An IoU-aware Dense Object Detector. *arXiv* **2020**, arXiv:2008.13367. [CrossRef]
64. Zhang, S.; Chi, C.; Yao, Y.; Lei, Z.; Li, S.Z. Bridging the Gap Between Anchor-based and Anchor-free Detection via Adaptive Training Sample Selection. *arXiv* **2019**, arXiv:1912.02424. [CrossRef]