

Dense Distinct Query for End-to-End Object Detection

Shilong Zhang^{1,3*}, Xinjiang Wang^{2*}, Jiaqi Wang⁴, Jiangmiao Pang⁴
Chengqi Lyu¹, Wenwei Zhang^{4,1}, Ping Luo^{3,1}, Kai Chen¹

¹Shanghai AI Laboratory²SenseTime Research

³The University of Hong Kong⁴S-Lab, Nanyang Technological University

Figure 1. Pros and Cons of different queries and their corresponding learning paradigms.

Abstract

One-to-one label assignment in object detection has successfully obviated the need for non-maximum suppression (NMS) as postprocessing and makes the pipeline end-to-end. However, it triggers a new dilemma as the widely used sparse queries cannot guarantee a high recall, while dense queries inevitably bring more similar queries and encounter optimization difficulties. As both sparse and dense queries are problematic, then what are the expected queries in end-to-end object detection? This paper shows that the solution should be Dense Distinct Queries (DDQ). Concretely, we first lay dense queries like traditional detectors and then select distinct ones for one-to-one assignments. DDQ blends the advantages of traditional and recent end-to-end detectors and significantly improves the performance of various detectors including FCN, R-CNN, and DETRs. Most impressively, DDQ-DETR achieves 52.1 AP on MSCOCO dataset within 12 epochs using a ResNet-50 backbone, outperforming all existing detectors in the same setting. DDQ also shares the benefit of end-to-end detectors in crowded scenes and achieves 93.8 AP on CrowdHuman. We hope DDQ can inspire researchers to consider the complementarity between traditional methods and end-to-end detectors. The source code can be found at <https://github.com/jshilong/DDQ>.

* Equal contribution.

1. Introduction

Object detection is one of the most fundamental tasks in computer vision, which aims at answering what objects are in an image and where they are. To achieve the objective, the detector is expected to detect all objects and mark each object with only one bounding box.

Due to the complex spatial distribution and the vast shape variance of objects, detecting all objects is quite challenging. To solve the problem, traditional detectors [17, 21, 27] first predefine dense grid queries to achieve a high recall. Convolutions with shared weights are then applied to quickly process dense queries in a sliding-window manner. At last, one ground truth bounding box is assigned to multiple similar candidate queries for optimization. However, the one-to-many assignment results in redundant predictions and thus requires extra duplicate-removal operations (e.g., non-maximum suppression) during inference, which causes misaligned inference with training and hinders the pipeline from being end-to-end (as shown in Fig. 1.(a)).

This paradigm is broken by DETR [2], which assigns only one positive query to each ground truth bounding box (one-to-one assignment) to achieve end-to-end. This scheme requires heavy computation to refine queries and adopts self-attention to model interactions between queries

¹ Anchors [17, 21] or anchor points [27] in conventional detectors play the same role as sparse object queries in [2]. Hence, we collectively refer to densely distributed anchor boxes and anchor points as dense queries.

to facilitate the optimization of one-to-one assignment, which unfortunately limits the number of queries. For example, DETR only initializes hundreds of learnable object queries. Therefore, compared to the densely distributed structures also struggle to distinguish between duplicate queries in conventional detectors, the sparse queries fall-cult optimization. In this study, DDQ FCN/R-CNN/DETR

Some recent works have also tried to integrate dense queries into one-to-one assignment [24, 28, 32]. However, dense queries in end-to-end detectors face unique challenges. For example, our analysis shows that this paradigm would inevitably introduce many similar queries (potentially representing the same instance) and that it suffers difficult and inefficient optimization as similar queries are assigned opposite labels under one-to-one assignment. (Fig. 1.(c)).

2. Related Work

Dense Queries with One-To-Many Assignment. One-stage detectors such as RetinaNet [17] and FCOS [27] use densely distributed queries for regression and classification. The same manner is also applied to the region proposal network (RPN) of multi-stage models [1, 21]. And one-to-many assignments are a common practice for these traditional detectors. Despite the fast development of one-to-many assignments from static label assignments (such as IOU-based [1, 17, 21] and center-based ones [13, 27]) to prediction-aware dynamic label assignments [5, 8, 9, 15, 35, 37], these strategies are also long criticized for they pair each ground truth with multiple queries and thus require additional postprocessing to remove duplicate predictions at inference, which prevents the pipeline from being end-to-end.

Sparse Queries with One-To-One Assignment. DETR [2] designs a small set of learned positional embeddings that represent the position in an image to focus on. These queries are then optimized with one-to-one assignments, making an end-to-end pipeline. Sparse R-CNN [26] reformulates queries in the traditional R-CNN framework as a bounding box and its corresponding embedding. Anchor DETR [31] provides the correspondence between anchor points and query position. DAB-DETR [19] explicitly learns a set of 4-D anchor boxes as queries. Though the formulation of queries varies, they share the same core idea of sparse queries and one-to-one assignments. Therefore, a low recall rate is an expected issue for these detectors.

Dense Queries with One-To-One Assignment. Both DeFCN [28] and OneNet [25] try to integrate one-to-one assignment with dense queries. Despite their competitive performance compared to FCOS [27], there is still a clear performance gap with recent detectors with dynamic one-to-many assignment strategies [8, 9, 12, 15, 37]. It is the optimization difficulty of similar queries under one-to-one assignments that accounts for the performance gap. Efficient DETR [32], and Two-Stage Deformable DETR [38] can also be regarded as a multi-stage version of this paradigm. Although DINO [33], Group DETR [4], and H-Deformable DETR [11] have introduced more positive sam-

Object detection in crowded scenes such as CrowdHuman [23] is another arena to testify to the effectiveness of DDQ. It is extremely cumbersome to tune the post-

3. Analysis of Sparse and Dense Queries

Current end-to-end detectors use either dense or sparse queries, both of which are however problematic during training. Specifically, sparse queries suffer a low recall rate, and dense queries have issues in optimization. To illustrate this, we increase the number of queries from 10 to 7000 in Sparse R-CNN, and the performance is shown as the black line in Fig. 2. The performance first keeps rising as the number of queries increases to around 2000, implying that the sparse queries (300) in Sparse R-CNN are far from enough due to the low recall rate. On the other hand, the performance finally plateaus and even decreases as queries number further increases. This phenomenon can be explained by the difficulty in distinguishing similar queries in end-to-end detectors with one-to-one assignment, especially when queries become denser.

at $0 < p < 0.5$ and may even cause negative training, (< 0) at $p > 0.5$.

As shown in the toy example, duplicated queries reduce gradients and even cause negative training, which dramatically suppresses convergence. To avoid this issue, we propose a distinct queries selection operation before the one-to-one assignment process. The distinct queries selection strategy is realized by a simple class-agnostic NMS. The filtered distinct queries are thus easier to optimize, and such an operation improves the performance by a clear margin, as seen from the red curve in Fig. 2. More surprisingly, the performance margin consistently increases along with more queries. A similar trend is also observed for Deformable DETR, which can be found in the supplementary material.

In other words, once we make sure the selected queries are distinct, the performance of Sparse R-CNN can be improved consistently with the increasing number of distinct queries. However, using a large number of distinct queries causes a significant memory footprint. For example, Sparse R-CNN requires around 45G memory per GPU with 7000 queries. To leverage the advantage of dense distinct queries (DDQ) with a reasonable computation cost, we give practical designs for all popular detector architectures (FCN, R-CNN, and DETRs).

4. Method

Dense distinct queries (DDQ) is our principle for designing an object detector and can be integrated into different architectures. We first briefly describe the design of DDQ followed by detailed descriptions for the three architectures: FCN, R-CNN, and DETRs. The overall pipeline is sketched in Fig. 3.

4.1. Paradigm of DDQ

Dense Queries As shown in Fig. 2, the memory cost soars for dense queries. The main reason for this is the heavy calculation for each query. Instead of adopting learnable positional embedding in DETR, DDQ directly takes the feature point on each feature map as densely distributed initial queries. The number of queries in the feature pyramid can easily surpass 10000 given an input resolution of 800x800. To discriminate dense queries with reasonable computation cost, a light-weighted convolutional/linear network serves as the first stage and processes all queries in a sliding window manner.

Distinct Queries Now that the importance of query distinctness for optimization has been revealed in Sec. 3, we would discuss in this section why a class-agnostic non-maximum suppression (NMS) can be used to select distinct queries and how it differs from the traditional NMS as post-processing in traditional detectors. Since each query represents a potential instance in an image, and an

Figure 2. The performance comparison of Sparse R-CNN with and without Distinct Queries Selection (a class-agnostic NMS with a threshold 0.7 before each region head). All models are trained using the standard 1x setting. The green dotted line represents the default number (300) of queries adopted in Sparse R-CNN. The subplot denotes the memory consumption per GPU as the number of queries increases in Sparse R-CNN.

To understand how similar queries would hinder optimization, we provide a simplified example where we assume there exist two identical queries. In this case, the one-to-one assignment assigns a foreground label to one of them but a background label to another. Without loss of generality, we adopt binary cross-entropy loss for classification. Therefore, the loss from these two queries becomes $L_1 = -\log(p_1) - \log(1 - p_2)$, where p_1 and p_2 are the probability scores of the positive and negative query, respectively, and satisfy $p_1 = p_2 = p$ as they are identical queries. In contrast, the loss value when only one of the duplicated queries exists is $L_0 = -\log(p)$. The ratio of the gradient between the duplicate and non-duplicate query is,

$$\frac{\partial L_1}{\partial p} = \frac{\partial L_0}{\partial p} = 1 - \frac{p}{1-p} \quad (1)$$

It is obvious that the gradient is scaled down (< 1)

Figure 3. The pipeline of DDQ. (a) shows the application of DDQ to an FCOS-like structure, which is a fully-convolutional network (FCN). It is thus dubbed DDQ FCN. The pyramid shuffle is applied to the last two and last convolution layers in the classification and regression branches, respectively. The class-agnostic NMS acts as the distinct queries selection operation. At last, only distinct queries will be assigned labels before calculating loss. (b) shows the design of DDQ for R-CNN structures (DDQ R-CNN). The last feature maps of the classification and regression branches of DDQ FCN are concatenated and iterated as distinct queries. The distinct queries are then sent to the refining heads with their corresponding bounding boxes. (c) shows the design of DDQ for DETRs (DDQ DETR). After distinct queries selection, the remaining feature embedding in the encoder is projected with a linear to the content part of distinct queries. Their corresponding bounding boxes will be mapped to the position embedding part. Both parts will be sent to 6 refining stages. In such a long refining architecture, DQS will be applied before each refining stage to ensure distinctness.

instance can be uniquely represented by its location in an efficient optimization in DDQ due to the removal of similar image [29], it comes naturally to detect similar queries. However, it also results in numerous "leaf" queries through using the class-agnostic overlapping ratio between them, which no gradients are back-propagated. Therefore, we debounding boxes predicted by queries. More specifically, sign an auxiliary head and an auxiliary loss to further harvest we apply a class-agnostic NMS to select distinct queries, harness the potential of the iterated queries following the design for the following one-to-one assignment. The loss is in DeFCN [28]. The auxiliary head is mostly identical to thus only computed on the selected distinct queries. As the main head, except that it adopts a soft one-to-many assignment should be noted that such an operation is adopted in both training and inference, instead of only in inference as an extra post-processing in traditional detectors. Therefore, can be found in our supplementary material.

such a pipeline still abides by the definition of end-to-end detectors. Compared to the training-unaware NMS in

traditional detectors, it is designed to relieve the burden of one-to-one assignment during training, and can thus be set to FCOS as an example of the FCN structure for object detection. It is found that dense queries are already available with an aggressive IoU threshold (0.7 in DDQ FCN and DDQ R-CNN, 0.8 in DDQ DETR), which is robust even on the dense feature pyramid. However, as the dense queries CrowdHuman dataset [23]. Such crowd scenes can not be processed level by level with convolutional layers. The properly handled by NMS as post-processing in traditional missing interaction across different levels poses a challenge detectors. We validate this in Table. 4. for the optimization of one-to-one assignments.

Loss Components

(1). Main Loss for Dense Distinct Queries We simply apply the bipartite matching algorithm in DETR [2] with the same cost weight in the one-to-one assignment. No extra prior (such as center priors in [28]) is adopted for a fair comparison with DETRs. After discriminating positive and negative samples, DDQ FCN adopts GIoU loss [22] and QFocal loss [16] with weights 2 and 1. For DDQ R-CNN and DDQ DETR, we just follow the implementation of Sparse R-CNN [26] and DINO [33].

(2). Auxiliary Loss for Dense Queries Despite the more

As shown in Fig. 3. (a), the DDQ principle is first applied to FCOS as an example of the FCN structure for object detection. It is found that dense queries are already available on the dense feature pyramid. However, as the dense queries are processed level by level with convolutional layers, the missing interaction across different levels poses a challenge for the optimization of one-to-one assignments.

Inspired by channel shuffle operation in ShuffleNet [36], we propose a pyramid shuffle to compensate for the interaction between queries in different levels where channels across adjacent levels are shuffled to form a new feature pyramid. Specifically, features at level i and level $i + 1$ are exchanged simultaneously. To account for the different spatial dimensions on the feature pyramid, a bilinear interpolation is adopted when exchanging features. We apply the pyramid shuffle operation on the last two and one convolution layer in the classification and regression branches, respectively. This approach stabilizes training and improves performance with negligible additional computation costs. In this work, we set

Figure 4. An illustration of pyramid shuffle. For queries in a specific scale level, it can do the interaction with queries in an adjacent level by shuffling channels. Before concatenating to the feature of the target level, the feature from other levels should be interpolate to the same size as the target level.

8 to 64 which means each feature level exchanges information from 128 channels with other levels. (Comparison with other approaches to model the interaction among dense queries, ablation, and the analysis of pyramid shuffle can be found in our supplementary material.)

As for the distinct queries selection module in DDQ FCN, we first select the top 1000 predictions according to the classification score from each feature level and then apply a class-agnostic non-maximum suppression with a threshold of 0.7 to ensure both distinctness and generality across different datasets.

4.3. DDQ R-CNN

We combine DDQ FCN with two refining stages in Sparse R-CNN to construct the DDQ R-CNN. As shown in Fig. 3 (b), thanks to the fast processing of dense distinct queries in DDQ FCN, we select 300 most representative queries according to the classification score from the remaining distinct queries. Then we concatenate the feature in the distinct position of the last feature map of the classification branch and regression branch to construct the query embedding. The query embedding and the corresponding bounding box prediction will be passed to the refinement head of Sparse R-CNN. Different from Sparse R-CNN which requires 6 stages of iterative query refinement, DDQ R-CNN needs as few as 2 refinement stages. Actually, the long iteration stages in Sparse R-CNN mainly compensate for the drawbacks caused by the sparse and sometimes similar input queries. For one thing, sparse queries could not cover all instances at initialization and thus need long cascading stages to refine. For another, similar queries also require long refinements to distinguish from each other to output a one-hot prediction for each instance [2]. In contrast, the dense distinct queries from DDQ R-CNN have addressed the above issues, and hence the number of iterative refinements can be significantly reduced. We also report the results when we change the number of queries and refinement heads of DDQ R-CNN in the supplementary material.

4.4. DDQ DETR

We construct DDQ DETR based on Deformable DETR*. As shown in Fig. 3 (c), We follow Two-Stage Deformable DETR [38] to process dense queries. Instead of initializing the content part with transformed coordinates, we fuse the feature map embedding of distinct positions as the content part, which makes the initial queries more distinct. A class-agnostic NMS with a threshold of 0.8 is set to select distinct queries before each refining stage. To compare with recent DETRs, we keep the original 6 refining stages and select K distinct queries for the refining stages. We also select the top K queries directly according to classification scores as dense queries for the auxiliary head in the decoder. The parallel forward of dense queries and distinct queries follows the H-Deformable DETR [11] and Group DETR [4]. We set K to 900, following DINO [33].

5. Experiments

In this section, we first introduce two standard benchmarks MS COCO [18] and CrowdHuman [23]. Then we introduce the setting of training and inference on both datasets. We also present three examples to show how end-to-end detectors with different architectures evolve to our DDQ step by step. At last, we compare DDQ with state-of-the-art conventional detectors and recent end-to-end detectors on MS COCO and CrowdHuman, which show that DDQ blends the advantages of two design paradigms. The latency of current popular models and DDQ is compared in our supplementary material.

5.1. Datasets

MS COCO 2017 [18] detection dataset is mainly used for comparison and ablation studies. It contains 118k training, 5k validation images, and 20k test images without annotations. There are on average 7 instances per image in this dataset. We report bounding box mean average precision (AP) as the performance metric, which is the mean average precision over multiple thresholds. If not specified, AP on the validation set is set as default.

Besides, we also report the performance on the CrowdHuman dataset [23], which has 15k training images and 4.4k validation images with around 23 heavily occluded instances per image. For evaluation, we use AP, mMR, and Recall as the metrics. mMR means the average log miss rate over false positives per image ranging from 10^{-2} to 10^0 following the official report [23]. A lower value of mMR means a better quality of high-scoring bounding boxes. All evaluation results are reported on the CrowdHuman validation subset.

* indicates it is an improved version based on techniques in DINO [33]. Details can be found in our supplementary material

5.2. Setting

COCO ResNet-50 [10] is the default backbone in this study if not specified. Most models adopt the 1x(12 epochs) training protocol in MMDetection [3]. AdamW [20] optimizer is used. For DDQ FCN, we set the initial learning rate to 5×10^{-5} and weight decay to 0.1. For DDQ R-CNN, we used a learning rate of 10^{-4} and weight decay of 0.05. The learning rate for both two CNN-based detectors decayed with a ratio of 0.1 at epoch 9 and epoch 12. For DDQ DETR, we utilized a learning rate of 10^{-4} and a weight decay of 0.05, and the learning rate decayed with a ratio of 0.1 at epoch 12 only. To ensure a fair comparison with other studies, we classified our data augmentation into three types: Normal, Multi-Scale, and DETR. Normal augmentation rescaled images to a short side of 800 pixels, with only random crops applied. For Multi-Scale augmentation, we used the classic multi-scale training range (480–800). Finally, the DETR augmentation followed that of the study by Carion et al. [2]. CrowdHuman ResNet-50 [10] is the default backbone. All conventional detectors and DDQ adopt the 3x (36 epochs) schedule with multi-scale training (480–800). All optimizer-related parameters are consistent with the setting on COCO. For end-to-end detectors with sparse queries, we follow the schedule (50 epochs) in Sparse R-CNN due to their slow convergence. The max detected instance number is changed to 500 for all conventional detectors following the [28]. For a fair comparison, We increase the number of queries to 500 for DDQ FCN/R-CNN, Sparse R-CNN, and Deform DETR. For DDQ DETR, we just keep the same 900 queries as COCO.

5.3. Evolving to DDQ

In this section, we show how detectors of different architectures evolve to DDQ. We can validate the importance of both density and distinctness from such a progressive development. From FCOS* to DDQ FCN In Table. 1, We start from an FCOS equipped with the bipartite matching algorithm in DETR [2] and our main loss components mentioned in Sec. 4.1, which is denoted as FCOS*. We adopt normal augmentation that is mentioned in 5.2 and train the model for 12 epochs. Due to the lack of cross-level interaction, its performance increases to 48.5 AP. However, initializing the performance is quite unstable and fluctuates between 24.5 AP and 36.5 AP in a few successful experiments. We select the best result 36.5 as our baseline. After adding pyramid shuffle operations to interact with cross-level queries, the training becomes stable and gets 1.1 AP improvement with only 0.2 G ops and 0.2 ms latency increase. Adding a distinct queries selection operation boosts the performance from 37.6 AP to 40.6 AP with only 0.3 ms latency. Such a 3 AP improvement demonstrates that the distinctness of queries is vital for the one-to-one assignment. After adding

an auxiliary loss for dense queries following DeFCN [28], we get DDQ FCN with a state-of-the-art performance of 41.5 AP. DQS on the strong baseline (equipped with pyramid shuffle and auxiliary loss) can be found in Table. 7. DQS still improves 2 AP (from 39.5 to 41.5).

Table 1. From FCOS* to DDQ FCN. PS stands for pyramid shuffle, and DQS means distinct queries selection operation. We also report the latency(L) and the ops(F).

Method	AP	AP ₅₀	AP ₇₅	L(ms)	F(G)
FCOS*	36.5	54.4	40.3	21.9	200.5
+ PS	37.6	56.3	41.3	22.1	200.7
+DQS	40.6	60.3	44.5	22.4	200.7
DDQ FCN	41.5	60.9	45.4	22.4	200.7

From Sparse R-CNN to DDQ R-CNN Table. 2 shows a progressive development from Sparse R-CNN to DDQ R-CNN. Sparse R-CNN with 300 queries achieves 39.4 AP within 12 epochs using the normal augmentation that is mentioned in Sec.5.2. Increasing the number of queries to 7000 improves the performance to 40.6 AP, at the cost of a quite heavy detector. Applying a distinct queries selection at the beginning of each stage boosts the performance by 2.5 AP to 43.1 AP. At last, by replacing the first four refinement stages with our DDQ FCN, which not only makes the input queries, the performance further increases to 44.6 AP.

Table 2. From Sparse R-CNN to DDQ R-CNN Q means the query, and DQS stands for the distinct queries selection

Method	AP	AP ₅₀	AP ₇₅	L(ms)	F(G)
Sparse R-CNN	39.4	57.7	42.5	31.0	160.2
+7000Q	40.6	58.7	44.0	135.0	781.0
+DQS	43.1	62.6	47.1	135.0	781.0
DDQ R-CNN	44.6	63.0	48.8	31.3	248.5

From Deformable DETR* to DDQ DETR Table 3 illustrates the progressive development from Deformable DETR* to DDQ DETR. Deformable DETR* achieves 45.4 AP with 900 queries within 12 epochs using the DETR augmentation mentioned in Sec.5.2. By employing a linear layer to process the dense queries on the feature pyramid and constructing content parts with feature embeddings, the content part as Two-Stage Deformable DETR(TS D-DETR) with mapped coordinates only achieves 46.7 AP, which is due to the lack of distinctness in the coordinates compared to the feature embedding. Adding an auxiliary loss for the decoder improves performance to 50.0 AP. Furthermore, by adding DQS before each refinement stage, the performance further increases to 50.7 AP. Finally, by adding the P2 feature and 100 CDN queries as in DINO [33], we achieve an impressive 52.1 AP, surpassing all detectors in the same setting. We show distinctness can be complementary to CDN

in Table. 7 and analyze the reason in our supplementary material to the dense queries that could cover most objects. For the other, DDQ also achieves the lowest mMR, as a merit of the distinctness among queries so that the detector can better differentiate false predictions.

Table 3. From Deformable DETR*(D-DETR) to DDQ DETR.

TS D-DETR stands for the naive two-stage version. Dense means initializing the content part with feature embedding. DQS stands for distinct queries selection. AUX-Decoder means the auxiliary loss for dense queries in the decoder. The ops only has comparative meaning and does not contain custom cuda operators

Method	AP	AP ₅₀	AP ₇₅	L(ms)	F(G)
D-DETR*	45.4	63.0	49.1	45	264
TS D-DETR	46.7	64.5	50.8	46	269
+Dense	48.5	66.2	52.7	47	270
+AUX-Decoder	50.0	67.4	54.8	47	270
+DQS	50.7	68.1	55.7	58	270
DDQ DETR _{scale}	52.1	68.9	57.3	114	860

5.4. Comparison with Other Detectors

Results on CrowdHuman We select some recent representative studies for comparison with DDQ on crowded scenes. It is seen that traditional detectors struggle between a low recall rate and a high false positive rate. Although DW [15] assignment is the recent state-of-the-art one-to-many assignment strategy and shows a clear increase in Recall compared to ATSS, it suffers from more serious false predictions and thus leads to a high mMR. The performance of such traditional detectors is limited by the post-processing NMS. In the supplementary material, we also show it can not be properly handled by adjusting the IoU threshold because it is training unaware.

End-to-end detectors can achieve a higher theoretical recall rate due to the removal of NMS as a post-process. However, a high recall is not guaranteed in Sparse R-CNN and Deformable DETR due to their sparse query design. Although DeFCN [28] achieves a better performance than other end-to-end methods by adopting dense queries, it is still difficult for DeFCN to distinguish between crowded objects and duplicated predictions (optimization difficulty) which affects the mMR.

Table 4. Performance on CrowdHuman

Method	Epochs	AP ₅₀	mMR	Recall
ATSS	36	89.6	44.4	95.9
DW	36	89.0	57.6	97.4
Cascade R-CNN	36	86.0	44.1	89.2
Sparse R-CNN	50	89.2	48.3	95.9
Deform DETR	50	89.1	50.0	95.3
DeFCN	36	91.0	46.5	97.9
DDQ FCN	36	92.7	41.0	98.2
DDQ R-CNN	36	93.5	40.4	98.6
DDQ DETR	36	93.8	39.7	98.7

In contrast, DDQ surpasses these detectors on all metrics by a clear margin. For one thing, DDQ leads in Recall due to suppress similar queries that slow the optimization. DDQ

Results on COCO We adopt heavier backbones and longer schedules to fairly compare with other detectors on COCO. As shown in Table. 5, we get all the results from the original study except those marked with *. We divide the results into two parts according to the augmentation. The first part adopts the augmentations in DETR [2] and reports the results on COCO validation dataset. DDQ remains its advantage among end-to-end object detectors using different backbone structures. It is worth emphasizing that DDQ FCN without any refinement architecture can already surpass most end-to-end detectors. DDQ R-CNN surpasses these methods by a large margin with only two refinement heads and without encoder architecture. The performance of DDQ R-CNN (R-50) can be further improved by adopting an encoder structure as in SEPC [30] or DyHead [6]. For example, It achieves an impressive 51.0 AP by adopting 6 blocks in DyHead as encoder structure, which is denoted as DDQ R-CNN_{with_encoder} (details about this model can be found in supplementary material). DDQ DETR outperforms recent DETR with a clear margin using R-50 as its backbone. When adopting a Swin-L backbone, it also surpasses the SOTA method DINO [33] by 0.7 AP. The second part adopts a multi-scale training (480-800) strategy for 24 epochs and reports the results on the COCO test-dev using ResNet-101, which is widely used by conventional detectors.

6. Ablation study

6.1. The Recall Improvement of Dense Queries

We analyze the recall of IoU threshold 0.5. As shown in Table. 6, we report the recall of the 5th stage input queries of Sparse R-CNN to make a fair comparison with the input queries of the refinement head in DDQ R-CNN. It can be seen that the Sparse R-CNN with 300 queries has a significantly lower recall (10.2 AP₉₀) than that with 7000 queries. In DDQ R-CNN, the queries from the DDQ FCN achieve a comparable recall to 7000 queries but with much less latency.

6.2. DQS with Different IoU Threshold

In this section, we show the robustness of distinct queries selection (DQS) with different IoU thresholds. As shown in Table. 7, the performance of DDQ FCN/R-CNN is quite robust when the IoU threshold ranges from 0.6 to 0.8. The performance drops slightly when the threshold is lower than 0.6, which is due to the lower recall rate for overlapping objects. The performance also starts to degrade when the

Table 5. Results on COCO Dataset. For DW, the * means we have retrained it with the same augmentation(480-800) as other methods using of cial implementation.

Method	Backbone	Val=Test	Epochs	AP	AP ₅₀	AP ₇₅	AP _s	AP _m	AP _l
Aug:DETR									
Cascade R-CNN [1]	ResNet-50	val	36	44.3	62.4	48	26.6	47.7	57.7
DAB DETR [19]	ResNet-50	val	50	42.6	63.2	45.6	21.8	46.2	61.1
DN-DETR [14]	ResNet-50	val	50	44.1	64.4	46.7	22.9	48.0	63.4
Deformable DETR [38]	ResNet-50	val	50	46.2	65.2	50.0	28.8	49.2	61.7
Efficient DETR [32]	ResNet-50	val	36	44.2	62.2	48.0	28.4	47.5	56.6
Sparse R-CNN [26]	ResNet-50	val	36	45.0	63.4	48.2	26.9	47.2	59.5
DINO _{4scales} [33]	ResNet-50	val	36	50.9	69.0	55.3	34.6	54.1	64.6
DINO _{5scales} [33]	ResNet-50	val	36	51.2	69.0	55.8	35.0	54.3	65.3
DDQ FCN	ResNet-50	val	36	44.8	64.1	49.4	29.9	47.8	56.0
DDQ R-CNN	ResNet-50	val	36	48.1	66.6	53.0	32.3	51.2	60.7
DDQ R-CNN _{with _encoder}	ResNet-50	val	36	51.0	69.0	56.0	34.0	54.4	64.6
DDQ DETR _{4scales}	ResNet-50	val	24	52.0	69.5	57.2	35.2	54.9	65.9
DDQ DETR _{5scales}	ResNet-50	val	24	52.8	69.9	58.1	37.4	55.7	66.0
Sparse R-CNN	ResNeXt-64x4d-101	test-dev	36	46.9	66.3	51.2	28.6	49.2	58.7
Deformable DETR	ResNeXt-64x4d-101	test-dev	50	49	68.5	53.2	29.7	51.7	62.8
DDQ FCN	ResNeXt-64x4d-101	test-dev	36	47.7	67.0	52.6	30.4	49.9	58.3
DDQ R-CNN	ResNeXt-64x4d-101	test-dev	36	49.9	68.8	54.8	31.8	52.2	61.7
Sparse R-CNN [26]	Swin-B	val	36	50.8	70.4	55.6	33.9	53.7	65.9
DDQ R-CNN	Swin-B	val	36	52.8	72.2	57.9	37.6	56.2	66.9
H-DeformableDETR _{4scales}	Swin-L	val	36	57.6	76.5	63.2	41.4	61.7	73.9
DINO _{4scales}	Swin-L	val	36	58.0	76.1	64.0	40.1	62.2	74.3
DDQ DETR _{4scales}	Swin-L	val	30	58.7	76.8	64.5	41.6	62.9	74.3
DDQ DETR _{4scales}	Swin-L	test-dev	30	58.8	77.0	64.6	39.4	62.1	74.0
Aug:Multi-Scale									
ATSS [34]	ResNet-101	test-dev	24	43.6	62.1	47.4	26.1	47.0	53.6
PAA [12]	ResNet-101	test-dev	24	44.8	63.3	48.7	26.5	48.8	56.3
OTA [9]	ResNet-101	test-dev	24	45.3	63.5	49.3	26.9	48.8	56.1
DW* [15]	ResNet-101	test-dev	24	45.8	64.6	49.6	27.3	48.9	57.0
DDQ FCN	ResNet-101	test-dev	24	45.9	65.1	50.7	28.3	48.6	55.6

Table 6. Recall improvement of dense distinct queries. Q means for the latency of the model.

Method	AR ₁₀₀	AR ₂₀₀	AR ₃₀₀	L(ms)
Sparse R-CNN	78.4	83.4	85.5	31.0
7000 Q & DQS	88.6	92.3	93.6	135.0
DDQ R-CNN	88.5	91.8	93.2	31.3

DETR exhibits a similar trend, as observed in Table 3. We can find even though CDN training in [33] has been adopted in DDQ DETR, distinctness still improves the performance. By the way, we also report the performance of ATSS [34] at various detectors including FCN, R-CNN, and DETRs. This different post-processing NMS IoU thresholds and show its behavior of DDQ in which a class-agnostic NMS is adopted can inspire researchers to consider the complementarity between traditional methods and end-to-end detectors.

7. Conclusion

This paper reveals that both sparse and dense queries in end-to-end detection are problematic. We propose that the expected queries should be both dense and distinct. Such

Table 7. Performance of DDQ on COCO when DQS adopts different IoU thresholds. Results of ATSS adopting different IoU thresholds in post-processing are also reported. None means we remove DQS or post-processing from the inference pipeline.* means the results are not stable and we report the average performance

COCO	0.5	0.6	0.7	0.8	0.9	None
DDQ FCN	40.8	41.4	41.5	41.4	40.5	39.5*
DDQ R-CNN	44.0	44.5	44.6	44.4	43.8	42.7*
DDQ DETR	50.1	50.7	50.9	51.3	51.0	50.7*
ATSS	39.3	39.5	39.3	38.7	36.7	19.6

8. Acknowledgement

This project is supported by the National Key R&D Program of China No.2022ZD0161600, No.2022ZD0161000 and the General Research Fund of HK No.17200622.

References

- [1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 2, 8
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *European Conference on Computer Vision* pages 213–229. Springer, 2020. 1, 2, 4, 5, 6, 7
- [3] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 6
- [4] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*, 2022. 2, 5
- [5] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. *arXiv preprint arXiv:2211.09788*, 2022. 2
- [6] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 7373–7382, 2021. 7
- [7] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2988–2997, 2021. 2
- [8] Chengjian Feng, Yujie Zhong, Yu Gao, Matthew R Scott, and Weilin Huang. Tood: Task-aligned one-stage object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* pages 3510–3519, 2021. 2
- [9] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 303–312, 2021. 2, 8
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition* pages 770–778, 2016. 6
- [11] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detsr with hybrid matching. *arXiv preprint arXiv:2207.13080*, 2022. 2, 5
- [12] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. *European Conference on Computer Vision* pages 355–371. Springer, 2020. 2, 8
- [13] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyond anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 2
- [14] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M. Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* pages 13619–13627, June 2022. 8
- [15] Shuai Li, Chenhang He, Ruihuang Li, and Lei Zhang. A dual weighting label assignment scheme for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 9387–9396, 2022. 2, 7, 8
- [16] Xiang Li, Wenhai Wang, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss v2: Learning reliable localization quality estimation for dense object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pages 11632–11641, 2021. 4
- [17] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 2
- [18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision* pages 740–755. Springer, 2014. 2, 5
- [19] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *International Conference on Learning Representation*, 2021. 2, 8
- [20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *International Conference on Learning Representation*, 2018. 6
- [21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99, 2015. 1, 2
- [22] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pages 658–666, 2019. 4
- [23] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 2, 4, 5
- [24] Peize Sun, Yi Jiang, Enze Xie, Wenqi Shao, Zehuan Yuan, Changhu Wang, and Ping Luo. What makes for end-to-end object detection? In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning* volume 139 of *Proceedings of Machine Learning Research* pages 9934–9944. PMLR, 18–24 Jul 2021. 2
- [25] Peize Sun, Yi Jiang, Enze Xie, Wenqi Shao, Zehuan Yuan, Changhu Wang, and Ping Luo. What makes for end-to-end object detection? *International Conference on Machine Learning* pages 9934–9944. PMLR, 2021. 2
- [26] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan

- Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14454–14463, 2021. 2, 4, 8
- [27] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 1, 2
- [28] Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with fully convolutional network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15849–15858, 2021. 2, 4, 6, 7
- [29] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. *European Conference on Computer Vision*, pages 649–665. Springer, 2020. 4
- [30] Xinjiang Wang, Shilong Zhang, Zhuoran Yu, Litong Feng, and Wayne Zhang. Scale-equalizing pyramid convolution for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) June 2020*. 7
- [31] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. *arXiv preprint arXiv:2109.07110*, 2021. 2
- [32] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021. 2, 8
- [33] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Harry Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 4, 5, 6, 7, 8
- [34] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 8
- [35] Xiaosong Zhang, Fang Wan, Chang Liu, Rongrong Ji, and Qixiang Ye. Freeanchor: Learning to match anchors for visual object detection. *Advances in neural information processing systems*, 32, 2019. 2
- [36] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856, 2018. 4
- [37] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020. 2
- [38] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 5, 8