Dense Distinct Query for End-to-End Object Detection

Shilong Zhang\textsuperscript{1,3,*}, Xinjiang Wang\textsuperscript{2,*}, Jiaqi Wang\textsuperscript{1}, Jiangmiao Pang\textsuperscript{1}, Chengqi Lyu\textsuperscript{1}, Wenwei Zhang\textsuperscript{4,1}, Ping Luo\textsuperscript{3,1}, Kai Chen\textsuperscript{1}

\textsuperscript{1}Shanghai AI Laboratory \hspace{0.5cm} \textsuperscript{2}SenseTime Research
\hspace{0.5cm} \textsuperscript{3}The University of Hong Kong \hspace{0.5cm} \textsuperscript{4}S-Lab, Nanyang Technological University

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 MS-COCO 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 \url{https://github.com/jshilong/DDQ}.

1. Introduction

Object detection is one of the most fundamental tasks in computer vision, which aims at answering \textit{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 \cite{17, 21, 27} first lay predefined dense grid queries\textsuperscript{1} 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 \cite{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, but

\begin{itemize}
  \item \textsuperscript{1} Anchors \cite{17, 21} or anchor points \cite{27} in conventional detectors play the same role as sparse object queries in \cite{2}. Hence, we collectively refer to densely distributed anchor boxes and anchor points as dense queries.
\end{itemize}
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 queries in conventional detectors, the sparse queries fall short in recall rate, as shown in Fig. 1.(b).

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)).

Now that both sparse queries (low recall) and dense queries (optimization difficulty) under one-to-one assignment are sub-optimal, what are the expected queries in end-to-end object detection?

In this study, we demonstrate that the solution should be dense distinct queries (DDQ), meaning that the queries for object detection should be both densely distributed to detect all objects and also distinct from each other to facilitate the optimization of one-to-one label assignment. Guided by such a principle, we consistently improve the performance of various detector architectures, including FCN, R-CNN, and DETRs. For one-stage detectors composed of fully convolutional networks (FCN), we first propose a pyramid shuffle operation to replace heavy self-attentions to model the interaction between dense queries. Then, a distinct queries selection scheme ensures that the one-to-one assignment is only imposed on the selected distinct queries, preventing contradictory labels from being assigned to similar queries. Such an end-to-end one-stage detector is named DDQ FCN and achieves state-of-the-art performance in one-stage detectors. DDQ also naturally extends to DETR and R-CNN structures [7, 19, 26, 38] by first laying dense queries as in [38] and then selecting distinct queries for later refining stages, which are respectively dubbed DDQ R-CNN and DDQ DETR.

We have conducted experiments on two datasets—MS-COCO [18] and CrowdHuman [23]. DDQ FCN and DDQ R-CNN obtain 41.5/44.6 AP, respectively, on the MS-COCO detection dataset [18] with the standard 1x schedule [17, 21]. Compared to recent DETRs, DDQ DETR achieved 52.1 AP in just 12 epochs with the DETR-style augmentation [2]. The strong performance demonstrates that DDQ overcomes the optimization difficulty in end-to-end detectors and converges as fast as traditional detectors with higher performance.

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-processing NMS in traditional detectors, as a low IoU threshold leads to missing overlapping objects, while a high threshold brings too many false positives. Recent end-to-end structures also struggle to distinguish between duplicated predictions and overlapping objects due to their difficult optimization. In this study, DDQ FCN/R-CNN/DETR achieve 92.7/93.5/93.8 AP and 98.2/98.6/98.7 recall on CrowdHuman [23], surpassing both traditional and end-to-end detectors by a large margin.

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, 12, 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 De-FCN [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-DeformableDETR [11] have introduced more positive samples to speed up convergence, the hindrance effect between similar queries and one-to-one assignments still remains unrevealed.

7330
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.

![Figure 2](image)

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 refining 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

\[
L = -\log(p_1) - \log(1 - p_2)
\]

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 \(\alpha\) of the gradient between the duplicate and non-duplicate query is

\[
\alpha = \frac{\partial L_1}{\partial p} / \frac{\partial L_0}{\partial p} = 1 - \frac{p}{1 - p}
\]

It is obvious that the gradient is scaled down (i.e., \(\alpha < 1\)) at \(0 < p < 0.5\) and may even cause negative training (i.e., \(\alpha < 0\)) 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 impose 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 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 act 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 filtered 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 refine stages. In such a long refine architecture, DQS will be applied before each refine stage to ensure distinctness.

instance can be uniquely represented by its location in an image [29], it comes naturally to detect similar queries using the class-agnostic overlapping ratio between the bounding boxes predicted by queries. More specifically, we apply a class-agnostic NMS to select distinct queries for the following one-to-one assignment. The loss is thus only computed on the selected distinct queries. It 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, 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 with an aggressive IoU threshold (0.7 in DDQ FCN and DDQ R-CNN, 0.8 in DDQ DETR), which is robust even on CrowdHuman dataset [23]. Such crowd scenes can not be properly handled by NMS as post-processing in traditional detectors. We validate this in Table. 4.

### 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 efficient optimization in DDQ due to the removal of similar queries, it also results in numerous "leaf" queries through which no gradients are back-propagated. Therefore, we design an auxiliary head and an auxiliary loss to further harness the potential of the filtered queries following the design in DeFCN [28]. The auxiliary head is mostly identical to the main head, except that it adopts a soft one-to-many assignment for dense queries to allow for dense gradients and more positive samples to speed up training. More details can be found in our supplementary material.

### 4.2. DDQ FCN

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 $S$ channels across adjacent levels are shuffled to form a new feature pyramid. Specifically, features at level $i$ exchange $S$ channels with those at level $i-1$ and level $i+1$ 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
For queries in a specific scale level, it can do the interaction with queries in an adjacent level by shuffling $S$ 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. $S$ 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 refine 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 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.

### 4.4. 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 $1.5K$ 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-DeformableDETR [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 in $[10^{-2}, 10^0]$ 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 $5 \times 10^{-5}$ 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 $2 \times 10^{-4}$ 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 flips 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 the 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 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 flops 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 flops (F).

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>AP_{50}</th>
<th>AP_{75}</th>
<th>L(ms)</th>
<th>F(G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FCOS*</td>
<td>36.5</td>
<td>54.4</td>
<td>40.3</td>
<td>21.9</td>
<td>200.5</td>
</tr>
<tr>
<td>+ PS</td>
<td>37.6</td>
<td>56.3</td>
<td>41.3</td>
<td>22.1</td>
<td>200.7</td>
</tr>
<tr>
<td>+ DQS</td>
<td>40.6</td>
<td>60.3</td>
<td>44.5</td>
<td>22.4</td>
<td>200.7</td>
</tr>
<tr>
<td>DDQ FCN</td>
<td>41.5</td>
<td>60.9</td>
<td>45.4</td>
<td>22.4</td>
<td>200.7</td>
</tr>
</tbody>
</table>

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 refining stages with our DDQ FCN, which not only makes the structure more light-weighted but also allows even denser 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>AP_{50}</th>
<th>AP_{75}</th>
<th>L(ms)</th>
<th>F(G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sparse R-CNN</td>
<td>39.4</td>
<td>57.7</td>
<td>42.5</td>
<td>31.0</td>
<td>160.2</td>
</tr>
<tr>
<td>+ 7000Q</td>
<td>40.6</td>
<td>58.7</td>
<td>44.0</td>
<td>135.0</td>
<td>781.0</td>
</tr>
<tr>
<td>+ DQS</td>
<td>43.1</td>
<td>62.6</td>
<td>47.1</td>
<td>135.0</td>
<td>781.0</td>
</tr>
<tr>
<td>DDQ R-CNN</td>
<td>44.6</td>
<td>63.0</td>
<td>48.8</td>
<td>31.3</td>
<td>248.5</td>
</tr>
</tbody>
</table>

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 performance increases to 48.5 AP. However, initializing 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 refining 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.

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

<table>
<thead>
<tr>
<th>Method</th>
<th>AP</th>
<th>AP_{50}</th>
<th>AP_{75}</th>
<th>L(ms)</th>
<th>F(G)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D-DETR*</td>
<td>45.4</td>
<td>63.0</td>
<td>49.1</td>
<td>45</td>
<td>264</td>
</tr>
<tr>
<td>TS D-DETR</td>
<td>46.7</td>
<td>64.5</td>
<td>50.8</td>
<td>46</td>
<td>269</td>
</tr>
<tr>
<td>+Dense</td>
<td>48.5</td>
<td>66.2</td>
<td>52.7</td>
<td>47</td>
<td>270</td>
</tr>
<tr>
<td>+AUX-Decoder</td>
<td>50.0</td>
<td>67.4</td>
<td>54.8</td>
<td>47</td>
<td>270</td>
</tr>
<tr>
<td>+DQS</td>
<td>50.7</td>
<td>68.1</td>
<td>55.7</td>
<td>58</td>
<td>270</td>
</tr>
<tr>
<td>DDQ DETR_{casefe}</td>
<td>52.1</td>
<td>68.9</td>
<td>57.3</td>
<td>114</td>
<td>860</td>
</tr>
</tbody>
</table>

5.4. Comparison with Other Detectors

Results on CrowdHuman We select some recent representational 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

<table>
<thead>
<tr>
<th>Method</th>
<th>Epochs</th>
<th>AP_{50}</th>
<th>mMR</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATSS</td>
<td>36</td>
<td>89.6</td>
<td>44.4</td>
<td>95.9</td>
</tr>
<tr>
<td>DW</td>
<td>36</td>
<td>89.0</td>
<td>57.6</td>
<td>97.4</td>
</tr>
<tr>
<td>Cascade R-CNN</td>
<td>36</td>
<td>86.0</td>
<td>44.1</td>
<td>89.2</td>
</tr>
<tr>
<td>Sparse R-CNN</td>
<td>50</td>
<td>89.2</td>
<td>48.3</td>
<td>95.9</td>
</tr>
<tr>
<td>Deform DETR</td>
<td>50</td>
<td>89.1</td>
<td>50.0</td>
<td>95.3</td>
</tr>
<tr>
<td>DeFCN</td>
<td>36</td>
<td>91.0</td>
<td>46.5</td>
<td>97.9</td>
</tr>
<tr>
<td>DDQ FCN</td>
<td>36</td>
<td>92.7</td>
<td>41.0</td>
<td>98.2</td>
</tr>
<tr>
<td>DDQ R-CNN</td>
<td>36</td>
<td>93.5</td>
<td>40.4</td>
<td>98.6</td>
</tr>
<tr>
<td>DDQ DETR</td>
<td>36</td>
<td>93.8</td>
<td>39.7</td>
<td>98.7</td>
</tr>
</tbody>
</table>

In contrast, DDQ surpasses these detectors on all metrics by a clear margin. For one thing, DDQ leads in Recall due 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.

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 refining 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 AR_{100}) 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 IoU threshold is larger than 0.8 due to its incapability to suppress similar queries that slow the optimization. DDQ...
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 a paradigm significantly improves the performance of various detectors including FCN, R-CNN, and DETRs. This proves the paradigm blends advantages from the traditional detectors and the recent end-to-end detectors. We hope it can inspire researchers to consider the complementarity between traditional methods and end-to-end detectors.

8. Acknowledgement

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


[26] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenguang Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan


