EASE-DETR: Easing the Competition among Object Queries

Yulu Gao[1,2] Yifan Sun[3] Xudong Ding[1,2] Chuyang Zhao[3] Si liu[1,2†]
[1]Institute of Artificial Intelligence, Beihang University
[2]Hangzhou Innovation Institute, Beihang University [3]Baidu VIS
{gyl97, zy2342107, liusi}@buaa.edu.cn {zhaochuyang,sunyifan01}@baidu.com

Abstract

This paper views the DETR's non-duplicate detection ability as a competition result among object queries. Around each object, there are usually multiple queries, within which only a single one can win the chance to become the final detection. Such a competition is hard: while some competing queries initially have very close prediction scores, their leading query has to dramatically enlarge its score superiority after several decoder layers. To help the leading query stands out, this paper proposes EASE-DETR, which eases the competition by introducing bias that favours the leading one. EASE-DETR is very simple: in every intermediate decoder layer, we identify the "leading / trailing" relationship between any two queries, and encode this binary relationship into the following decoder layer to amplify the superiority of the leading one. More concretely, the leading query is to be protected from mutual query suppression in the self-attention layer and encouraged to absorb more object features in the cross-attention layer, therefore accelerating to win. Experimental results show that EASE-DETR brings consistent and remarkable improvement to various DETRs.

1. Introduction

The non-duplicate detection ability is an important characteristic shared by DEtection Transformer (DETR) [3] and its variants [6, 18, 21, 26, 30, 34, 35, 38, 39, 42]. More concretely, DETR uses plenty of queries (*e.g.*, 300 in DETR [3] and 900 in DINO [39]) in the decoder to search for the objects. Though the queries are redundant, the predictions made by them are expected to be non-duplicate, *i.e.*, only one query predicts each ground-truth object. This characteristic is known as being related to the one-to-one label assignment, which sets up the training objective. In contrast, we are interested in the mechanism DETR employs to reach this objective.

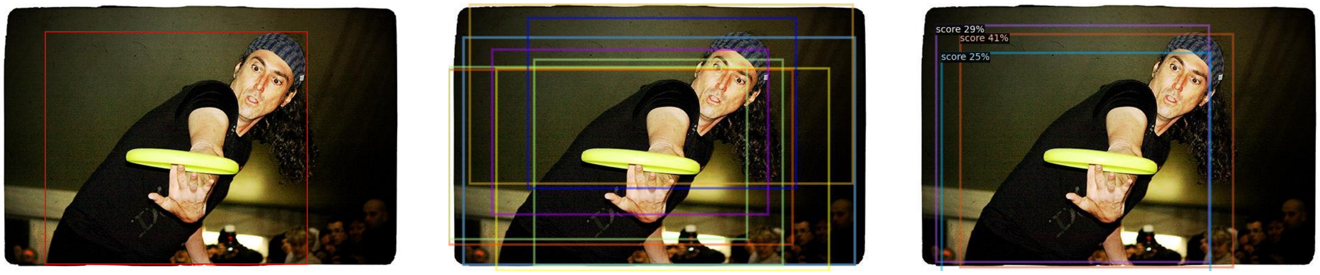
This paper views the non-duplicate detection ability of DETR as the result of a competition among object queries. We explain this viewpoint through a revisit into the DETR decoder. Before the first decoder layer, there are multiple queries around each object, as shown in Figure 1. Some queries become even closer after the first decoder, because they are attracted by the same object through cross-attention. Such closeness regards not only the position, but also the predicted scores. Within these close queries, only a single one can win the chance of detecting the object and makes high prediction score at the final decoder layer. All the other competing queries are suppressed as the background and make low prediction scores. Therefore, a leading query (that usually has only subtle superiority at the beginning) has to dramatically enlarge its score gap against the trailing queries after several decoder layers.

We argue that helping the leading query to win the above competition easier can benefit the training efficiency and improve the detection accuracy. To achieve this, we propose EASE-DETR, which eases the competition by introducing bias that favours the leading query. EASE-DETR is very simple: given any two competing queries in the intermediate decoder layers, we identify their relationship of "leading / trailing" and further amplify the superiority of the leading query in the following layer. The superiority amplification can be conducted in both the following *self-attention* and *cross-attention* layers.

- *The self-attention layer* facilitates global interaction among all queries for suppressing duplicate detections [14, 26]. However, any two queries in the self-attention are in symmetric position, and both receive suppression from each other. We aim to decay the suppression upon the leading query while maintaining the suppression upon the trailing queries, so as to enlarge their gap. Instead of hand-crafted decay, we encode the binary "leading / trailing" relationship into a decay weight through a trainable projection (MLP, in practice). Empirically, we find the learned decay weights protect the leading query from mutual suppression and thus enlarge its superiority against the trailing one.

- *The cross-attention layer* enables each query to absorb

[†] Corresponding author.



(a) ground truth (b) before decoder (c) after first decoder layer

Figure 1. DETR decoder has multiple queries competing for each ground-truth object. These competing queries have close position and similar prediction scores at their initial status. (a) the ground truth; (b) multiple queries are around the same object before input into the decoder; (c) after the first decoder, some queries are attracted by the object through cross-attention. Correspondingly, they become even closer and have close prediction scores.

features. Particularly, the competing queries absorb object features from the same object and thus tend to make duplicate predictions. To ease the competition in the cross-attention layer, we aim to decay the feature absorption of the trailing queries while maintaining that of the leading query. We use another trainable projection to encode the binary relationship into a decay weight, which enforce stronger decay upon the trailing queries. Consequently, the leading query increases its priority for learning object feature and enlarges its superiority.

In both the self-attention layer and cross-attention layer, the operation of EASE-DETR is very simple and similar, *i.e.*, multiplying the original attention score with an additional decay weight projected from the binary "leading / trailing" relationship. It adds virtually no computational cost to training and inference. In practice, we develop another variant that combines the spatial relationship (*i.e.* the IoU) with the score relationship for generating the decay weight, because we want the impact of easing competition is to be within local regions. This variant brings further improvement. No matter using IoU or not, there is no hand-crafted hyper-parameter. All additional projections are learned in the end-to-end DETR training.

We conduct extensive experiments with popular DETR baselines on MS-COCO [19] dataset. Experimental results show that our EASE-DETR brings remarkable improvement to DETRs. For example, on the Deformable++ DETR [16] baseline (ResNet-50 [13] backbone), EASE-DETR achieves 1.3 AP improvement under the 12-epoch training scheme. Importantly, EASE-DETR, with focus on competition under the one-to-one supervision, manifest good cooperation with the one-to-many supervision strategy: the one-to-one branch uses the EASE-DETR attention while the one-to-many branch uses the standard attention. Such a simple combination achieves competitive accuracy, *e.g.*, 50.8 AP on ResNet-50 [13] DINO [39] baseline and 57.8 AP on Swin-Large [25] DINO [39].

The contributions of our work are summarized as follows: First, we conduct an in-depth investigation of the DETR decoder, uncovering how attention weights influence the suppression process of similar queries. Second, we introduce explicit relationships between queries into the decoder, proposing the MSelf-attention layer and MCross-attention layers to enhance the suppression process of queries. Third, through empirical experiments, we demonstrate that EASE-DETR achieves performance improvements across various popular DETR baselines and ultimately surpasses the state-of-the-art results.

2. Related Work

2.1. DETR-base Detectors

The landscape of object detection has been fundamentally altered with the advent of the original DEtection TRansformer (DETR) [3]. Introduced as a pioneering approach, DETR offered an end-to-end framework that mitigated the reliance on hand-crafted components prevalent in previous methodologies, such as Non-maximum Suppression (NMS). While DETR set a new paradigm, it still suffers from its comparatively slow training convergence and sub-optimal performance on detecting smaller objects. To solve these problems, a series of innovative works have emerged. Deformable DETR [42] introduces multi-scale features and thus proposes a deformable attention module to detect small objects and speed up training. Further adaptations, such as the DAB-DETR [21], enhance the DETR's performance by integrating anchor boxes into queries explicitly and update them continuously. DN-DETR [18] introduces denoising part to stabilize the bipartial matching, and DINO [39] and its successor Stable DINO [22], offer refined optimization techniques that stabilize the training process and enhance the convergence speed. The recent DAC-DETR [14] divides the cross-attention out from this contrary for better conquering thus accelerates convergence and improves its performance. In this paper, we introduce a plugin solution to ease the query competition in DETR-based detectors, offering enhanced compatibility with various detector architectures.

2.2. Modulated Attention

The transformer [33] revolutionizes deep learning by using attention mechanisms to efficiently blend queries and keys. And DETR [3] uses transformer structure on object detection, results an end-to-end paradigm on object detection. Deformable DETR [42] mudulates traditional attention with deformable attention for sampling efficiency and speed. Conditional DETR [26] uses object query conditioning to hasten convergence. SMCA [10] proposes Spatially Modulated Co-Attention mechanism which modulates co-attention emphasizing relevant image regions for quick training. DAB-DETR [21] modulates the queries with dynamic anchor boxes to improve the query-to-feature similarity and eliminate the slow training convergence issue in DETR. DDQ [40] utilizes NMS to select query for both training and inference. Some works [8, 9, 11, 15, 31, 37] modify attention mechanisms in other fields. The aforementioned research modifies the traditional DETR model to cater to specific issues, achieving notable performance improvements. Our focus lies in the relationships between each query. We encode these relationships into the self-attention and cross-attention of the decoder to further boost performance.

2.3. Non-duplicate Detection

The traditional detectors [1, 4, 12, 17, 20, 23, 28, 29, 32, 36] employ Non-Maximum Suppression (NMS) [27] to eliminate redundant detections, ensuring unique and precise detection results. Alternatively, CenterNet [41] approaches duplicate results problem through NMS Pooling, which centralizes on the object's central point rather than bounding boxes, pooling nearby detection responses, and effectively managing dense object scenarios. Moreover, DETR [3] and its variants employ self-attention layers for information propagation, combined with a one-to-one label assignment strategy to address the issue of duplicate detection results. This paper proposes a method to ease query competition, a challenge intrinsic to these models, aiming to achieve superior non-duplicate detection results.

3. Methodology

3.1. Overview

In our EASE-DETR development, we introduced a novel approach to ease query competition by introducing a bias towards the leading query, thereby enhancing model convergence and accuracy. Our architecture, depicted in Figure 2 (A), refines the traditional DETR framework, particularly in the decoder's self-attention and cross-attention modules. We start by computing scores and bounding boxes for each query in the decoder, using these scores for calculate relative ranking relation R_{rank} and the bounding boxes to calculate spatial relation R_{spt}. And the information is then encoded into the MSelf-Attention mechanism, adjusting attention weights to emphasize the leading query. Concurrently, in the MCross-Attention mechanism, we assess and scale the attention based on the R_{rank} and R_{spt}. These enhancements are parallelized and minimally increase the number of parameters, ensuring EASE-DETR's compatibility with mainstream DETR detectors and enhancing performance without adding computational overhead.

3.2. Preparation

In the DETR, each intermediate transformer decoder layer predicts scores and boxes for auxiliary supervision. These outputs are precisely the inputs required by our EASE model. We predict the score and bounding box from the preceding layer, utilizing a category-independent score. These scores are employed to establish a 'leading/trailing' relationship among pairs of queries. The formula for the relative ranking R_{rank} is as follows, where S_i, S_j represents the prediction score of the i-th query from the previous layer:

$$R_{rank}(i,j) = \begin{cases} +1 & S_i \geq S_j \\ -1 & S_i < S_j \end{cases}. \tag{1}$$

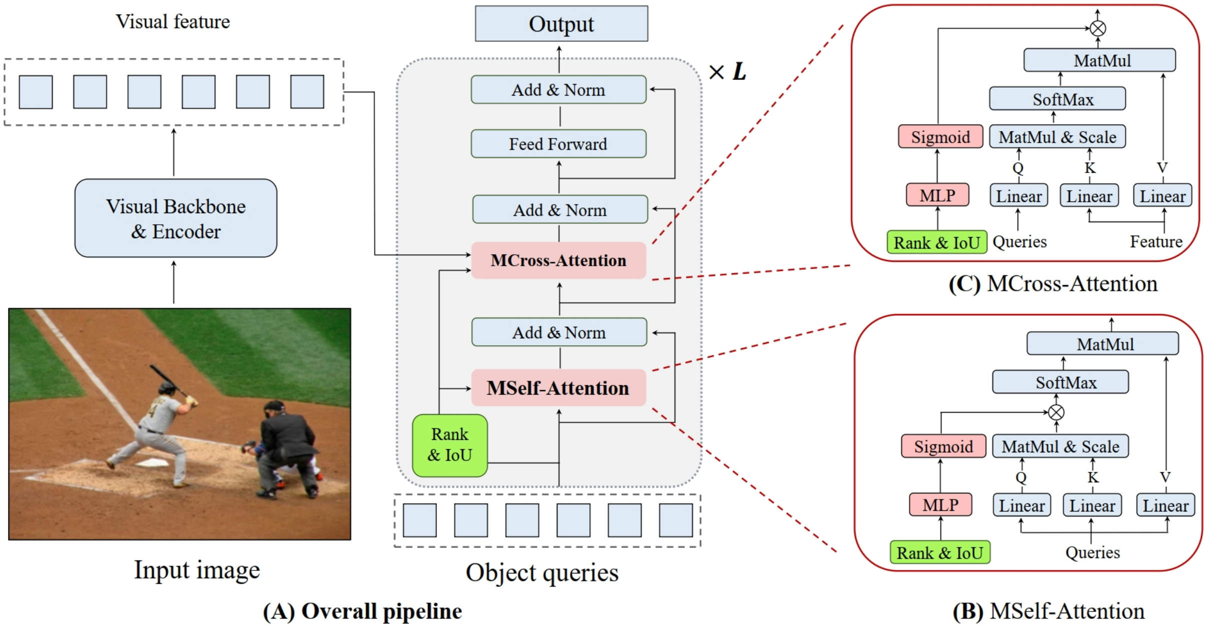
Our objective with the relative ranking relationship R_{rank} is to ease the competition among similar queries. To further strengthen the aspect of locality in our methodology, we introduce the spatial relation R_{spt} between queries into our later computations. The formula is as follows:

$$R_{spt}(i,j) = IOU(B_i, B_j). \tag{2}$$

The relative ranking relation, denoted as R_{rank}, and the spatial relation, denoted as R_{spt}, will be utilized to modulate the subsequent self-attention and cross-attention module.The bounding boxes B_i, B_j are produced from the previous layer.In the first layer, we bypass the use of the EASE module and adopt the same box initialization as the baseline.

3.3. Encourage Leading Query in Self-attention

In our analysis of the DETR framework, we emphasize the critical role of self-attention in facilitating non-duplicated detection. And the interaction of self-attention mechanism introduce competition among queries. These interactions occur as each query is projected and subsequently undergoes matrix multiplication to calculate attention scores. At this stage, the presence of competing queries with very similar initial prediction scores necessitates extensive learning for effective optimization. To address this, our method



(A) Overall pipeline

(B) MSelf-Attention

(C) MCross-Attention

Figure 2. **(A)** EASE-DETR uses the relative ranking ("leading / trailing") and IoU between every two queries to ease their competition. EASE-DETR is very simple and makes only slight modifications to the original self-attention and cross-attention layers. All the modifications are highlighted in red. **(B)** In the modified self-attention layer (MSelf-Attention), the ranking and IoU are multiplied and then encoded through an MLP and sigmoid function. Through end-to-end training, it learns to decay the suppression upon the leading queries. **(C)** In the modified cross-attention layer (MCross-Attention), the ranking and IoU are multiplied and then encoded into another decay weight. This decay weight reduces the object feature absorbed into trailing queries. In a word, both the MSelf-Attention and MCross-Attention layers introduce bias that favours the leading query and thus help it to enlarge its superiority in the competition.

involves binarizing the relative ranking R_{rank}, encoding them, and subsequently utilizing this encoded data to modulate the attention weights.

A straightforward approach to modulate attention using relative ranking involves directly utilizing R_{rank} as the sign bit for the attention weights after softmax operation. This method is notably simple, and we have corroborated its effectiveness in our subsequent experiments, which encompass both 2D and 3D detection tasks.

To strengthen the aspect of locality, we further integrate R_{rank} and R_{spt} in our approach. The R_{rank}, being binary, is denoted by $+1$ and -1. We directly multiply the R_{spt} values with R_{rank} and use MLP to binarize them.The MLP below uses a single hidden layer with a dimension of 16. The corresponding formula is as follows:

$$D^s = \texttt{sigmoid}(\texttt{MLP}(R_{rank} \cdot R_{spt})). \tag{3}$$

The resulting self attention decay D^s modulate the attention in the self-attention mechanism, as illustrated by the following equation:

$$A_{ij} = \frac{D^s_{ij} \exp(W_{ij})}{\sum_{k=1}^{N} D^s_{ik} \exp(W_{ik})}, \tag{4}$$

where W represent original attention weight before softmax, and A represents the final attention weight. In this formula, we modulate the attention scheme by multiplying the decay D^s with the respective attention weights. This modulation either intensifies or diminishes the original attention signals, thereby embedding explicit query relationships into the model.

In our methodology, a single decay matrix D^s is shared across all attention heads. To accommodate the multi-head attention architecture, we have refined the output layer of the MLP, transitioning from a unidimensional output to an N-dimensional output, where N stands for the number of heads in the self-attention mechanism. With this modification, each head is allocated a corresponding decay matrix D^s, which is uniquely modulated to fine-tune the processing capabilities of that particular head.

3.4. Penalize Trailing Query in Cross-attention

After modulating self-attention with R_{rank} and R_{spt}, we further explored their modulation effects on cross-attention. Typically, cross-attention is considered a mechanism for aggregating object information. Our aim in the cross-attention phase is to encourage the leading query to assimilate more

Method	Backbone	#epochs	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
Deformable++ [16]	R50	12	47.0	65.2	51.5	31.2	50.4	61.1
Deformable++ [16]	R50	36	49.0	67.6	53.5	32.6	52.3	63.3
Deformable++ [16]	Swin-T	12	49.3	–	–	31.6	52.4	64.6
EASE-Deformable++ (ours)	R50	12	48.3 (+1.3)	67.0	52.3	31.5	51.3	62.8
EASE-Deformable++ (ours)*	R50	24	49.6 (+0.6)	68.4	54.1	32.8	52.8	64.2
EASE-Deformable++ (ours)	Swin-T	12	50.2 (+0.9)	69.4	54.8	32.5	53.6	65.3
H-DETR [16]	R50	12	48.7	66.4	52.9	31.2	51.5	63.5
H-DETR [16]	Swin-T	12	50.6	-	-	33.4	53.7	65.9
EASE-H-DETR (ours)	R50	12	49.4 (+0.7)	67.7	53.8	33.6	52.7	63.6
EASE-H-DETR (ours)	Swin-T	12	51.7 (+1.1)	70.2	56.3	34.5	55.0	66.8
DINO [39]	R50	12	49.0	66.6	53.5	32.0	52.3	63.0
DINO [39]	Swin-T	12	51.3	69.0	56.0	34.4	54.4	66.1
EASE-DINO (ours)	R50	12	49.7 (+0.7)	67.5	54.3	32.7	52.9	64.1
EASE-DINO (ours)	Swin-T	12	51.8 (+0.5)	69.5	56.6	35.4	55.1	66.1

Table 1. EASE-DETR brings consistent improvement over popular baselines. The term 'EASE-' indicates the integration of our EASE module. We rigorously test its performance with different detectors, over various epochs, and using diverse backbones. The integration of our EASE module consistently results in performance enhancements. Notably, in the 12-epoch setting with ResNet-50 backbone, significant improvements are observed in the currently popular Deformable++ and DINO frameworks. Specifically, there is an increase of 1.3 AP in Deformable++ [16] and a 0.7 AP increase in DINO [39].

Method	Backbone	#epochs	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
Baseline (DINO [39])	R50	12	49.0	66.6	53.5	32.0	52.3	63.0
Baseline (DINO [39])	R50	24	50.4	68.3	54.8	33.3	53.7	64.8
H-DETR [16]	R50	12	48.7	66.4	52.9	31.2	51.5	63.5
H-DETR [16]	R50	36	50.0	68.3	54.4	32.9	52.7	65.3
Group-DETR [5]	R50	12	49.8	-	-	32.4	53.0	64.2
Stable-DINO-4scale [22]	R50	12	50.4	67.4	55.0	32.9	54.0	65.5
Stable-DINO-4scale [22]	R50	24	51.5	68.5	56.3	35.2	54.7	66.5
DAC-DETR [14]	R50	12	50.0	67.6	54.7	32.9	53.1	64.2
DAC-DETR [14]	R50	24	51.2	68.9	56.0	34.0	54.6	65.4
EASE-DETR (ours)	R50	12	50.8 (+1.8)	68.9	55.3	34.1	54.2	65.1
EASE-DETR (ours)	R50	24	51.6 (+1.2)	69.9	56.2	34.0	54.6	66.0
Baseline (DINO [39])	Swin-L	12	56.8	75.6	62.0	40.0	60.5	73.2
Group-DETR [5]	Swin-L	36	58.4	-	-	41.0	62.5	73.9
Stable-DINO-4scale [22]	Swin-L	12	57.7	75.7	63.4	39.8	62.0	74.7
Stable-DINO-4scale [22]	Swin-L	24	58.6	76.7	64.1	41.8	63.0	74.7
DAC-DETR [14]	Swin-L	12	57.3	75.7	62.7	40.1	61.5	74.4
EASE-DETR (ours)	Swin-L	12	57.8 (+1.0)	76.7	63.3	40.7	61.9	73.7

Table 2. Comparison to SOTA DETR. Compared to DINO [39], our EASE-DETR yields improvements of 1.8 and 1.2 AP in 12-epoch and 24-epoch settings on the ResNet-50 [13] backbone, achieving 50.8 and 51.6 AP, respectively. The results surpasses methods like Stable-DINO [22], Align-DETR [2], and DAC-DETR [14]. Additionally, using the Swin-L backbone with EASE-DETR increases performance by 1.0 AP, reaching 57.8 AP and exceeding the latest SOTA, Stable-DINO.

object features in the cross-attention layer, while diminishing the ability of trailing queries to acquire target information. To achieve this, we adopted a straightforward approach: we quantified the level of suppression L using the product of $-R_{rank} \cdot R_{spt}$, as illustrated by the following equation:

$$L_i = \max_{j=0}^{n-1}(-R_{rank}(i,j) \cdot R_{spt}(i,j)), \tag{5}$$

where the n index number of query. And the specific formula for cross attention decay D^c is as follows,

$$D_i^c = \texttt{sigmoid}(\texttt{MLP}(L_i)). \tag{6}$$

For the i-th query, our methodology involves selecting queries that not only have a higher score than the current query but also share an IoU with it. We then compute the cumulative attention weight of these queries in relation to the current query. This sum is subtracted from one to derive the cross-attention scale, denoted as D_i^c. We employ D_i^c to modulate the result of the current cross-attention directly by multiplying it with D_i^c. Similarly, drawing inspiration from the multi-head design inherent in self-attention mechanisms, our model generates a multi-head scale D_i^c within the cross-attention.

4. Experiments

4.1. Setup and Implementation Details

Dataset. We conduct experiments on the COCO [19] 2017 validation dataset. And we report our results with the mean average precision (mAP) metric under different IoU thresholds. We report results with two different backbones: ResNet-50 [13] (pretrained on ImageNet-1k [7]) and Swin-L [25] (pretrained on ImageNet-22k [7]).

Implementation details. We develop EASE-DETR using Python 3.8.17, PyTorch 1.10.1, and CUDA 11.3. For 2D object detection training, we train all models with a batch size of 16, employing the AdamW optimizer. We initialized the learning rate at 1e-4. In the setting with 12 epochs, we reduced it by a factor of 0.1 at the 11th epoch. Similarly, in the setting with 24 epochs, we decreased the learning rate by a factor of 0.1 at the 20th epoch. In our experiments on 3D object detection, we adhere to the established configurations from PETR [24], adapting only the R_{rank} component for the 3D detector.

4.2. Main Results

In our experiments conducted on the COCO 2017 validation set, we evaluate the effectiveness of our EASE module. The results, as presented in Table 1, demonstrate the module's performance across various detectors, epochs, and diverse backbone architectures. The integration of our EASE module consistently leads to enhancements in performance. Notably, when incorporated into the original attention architecture, the module enables a significant improvement in Deformable++ models, achieving an increase of 1.3 AP and 0.9 AP on the ResNet50 and Swin-T backbones, respectively. Remarkably, our EASE-enhanced Deformable++ outperforms the standard Deformable++ by 0.6 AP in just 24 epochs, compared to the latter's 36 epochs.

Furthermore, applying EASE to H-DETR [16] and DINO [39] models also yields substantial improvements, signifying the module's compatibility and efficacy in one-to-many and denoise design paradigms. These results underscore the adaptability and effectiveness of our EASE module in various frameworks and settings.

To compare with current SOTA DETR models, we utilize DINO [39] as our baseline, integrate our EASE module, and adopt the recent one-to-many strategy from DAC-DETR [14], a method we refer to as EASE-DETR, as showcased in Table 2. Remarkably, this integration yields improvements of 1.8 and 1.2 AP for the 12-epoch and 24-epoch settings on the ResNet-50 [13] backbone, respectively, achieving 50.8 AP and 51.6 AP. These results surpass those of other leading methods such as Stable-DINO[22], Align-DETR[2], and DAC-DETR[14]. In a thorough evaluation, we experiment with EASE-DETR using the Swin-L backbone, further boosting performance with an additional increase of 1.0 AP, reaching a new high of 57.8 AP, and surpassing the recent SOTA model, Stable-DINO[22].

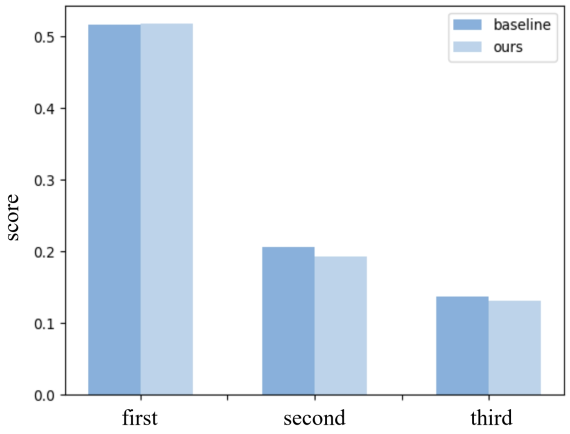


Figure 3. Comparison of the predicted score distribution of the competing queries in the last decoder layer. For each object, the top-scored query wins the competition and becomes the final prediction, while the other queries are suppressed to be background and should have low scores. 'First', 'Second', 'Third' represent the queries with the 1st, 2nd and 3rd highest prediction score, respectively. It shows that our EASE-DETR increases/decreases prediction score for the 1st/2nd queries, therefore enlarging their gap and easing the query competition.

Furthermore, we compare the predicted score distribution of the competing queries in the last decoder layer, illustrated in Figure 3. For each object, the top-scored query wins the competition and becomes the final prediction, while the other queries are suppressed to be background and should have low scores. On the val set, EASE-DETR enlarges the gap between the wining query and the competitors, indicating eased competition.

4.3. Ablation Studies

Component analysis. We conduct experiments to ascertain the effectiveness of each module within our proposed framework. The detailed outcomes of these experiments are presented in Table 3. Our study commences with a baseline model employing the Deformable DETR architecture. The model exhibits continuous improvements through the sequential addition of MSelf-attention and MCross-attention modules. This progression significantly highlights the contribution of each module to enhancing the overall perfor-

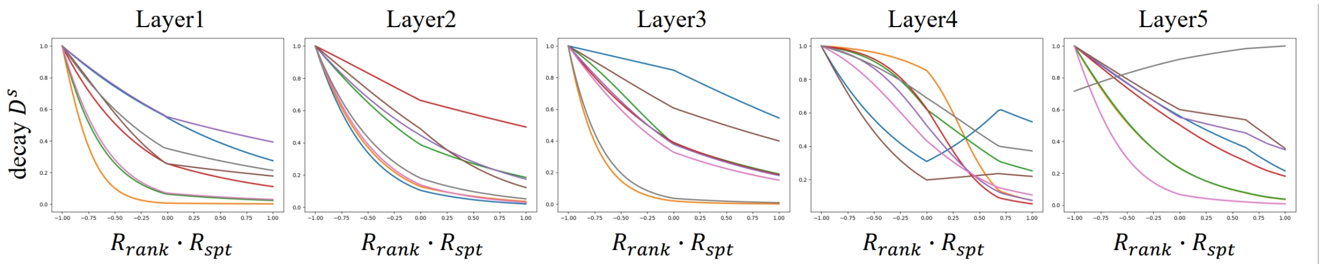


Figure 4. Visualizing the relation between $R_{rank} \cdot R_{spt}$ and self attention decay D^s. The graph reveals that, across different layers and heads, the relationship between $R_{rank} \cdot R_{spt}$ and D^s generally displays a monotonic decrease, with the function being asymmetric, thus underscoring the importance of introducing R_{rank}.

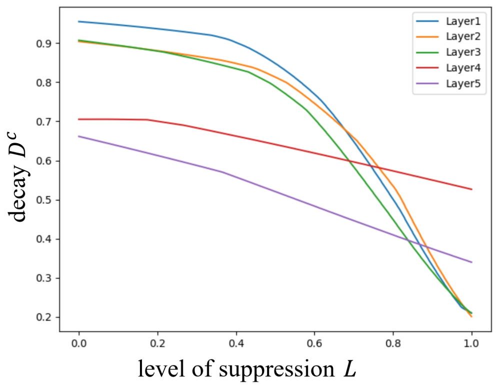


Figure 5. Visualizing the relation between level of suppression L and cross attention decay D^c. The illustration reveals that with an increasing value of L, the decay D^c diminishes, leading to a weaker absorption of target information. Furthermore, it is observed that the initial layers exhibit a stronger absorption of object information compared to the later layers.

mance of the model. Specifically, the integration of MSelf-attention and MCross-attention modules leads to incremental improvements of 1.1 AP and 0.2 AP, respectively. Notably, our model achieves an impressive 48.3 AP on the Deformable++ baseline, even without the implementation of a one-to-many approach.

We visualize the relation between $R_{rank} \cdot R_{spt}$ and self-attention decay D^s as shown in Figure 4, where each self-attention layer includes 8 heads. The graph reveals that, across different layers and heads, the relationship between $R_{rank} \cdot R_{spt}$ and D^s generally displays a monotonic decrease, with the function being asymmetric, thus underscoring the importance of introducing R_{rank}. This observation is consistent with our analysis of attention modulation. Additionally, we visualize the relation between the level of suppression L and cross-attention decay D^c in Figure 5. It is apparent that a higher L corresponds to a smaller decay value, indicating that stronger suppression of the query leads to less absorption of object information. Moreover, the initial layers show a greater absorption of object information compared to the later layers.

Comparison of different self-attention modulation strategies. In our modulated self-attention study, we introduced two variant approaches to ease query competition by favoring the leading query. In the first variant, we solely incorporated relative ranking R_{rank} to modulate self-attention. This method acts as a sign factor, altering attention post-softmax, as shown in the 2nd row of Table 4. Compared to the baseline, introducing only R_{rank} leads to an improvement of 0.8 AP. The second variant aims to enhance locality by building upon R_{rank} with the addition of R_{spt}. Considering that R_{rank} is binary and R_{spt} is a non-negative number, we multiply the two as input. Through a MLP and sigmoid, we generated weights to adjust the self-attention weights. We perform the modulation of our second method variant before the softmax function, correspond to the 3rd row. The results further improved by 0.3 AP on top of the first variant, culminating in a total enhancement of 1.1 AP compared to the baseline.

Investigation MSelf-attention on two-stage Deformable DETR. In order to delve deeper into the MSelf-attention module, we conducted ablation studies on commonly used tricks, as demonstrated in Table 5. It is evident from the results that our Mself-attention consistently enhances performance across these experiments, validating the effectiveness of our approach. Notably, integrating MQS (Mixed Query Selection) on top of our method yields a relatively modest improvement (from 44.3 AP to 46.3 AP compared to 47.4 AP to 47.7 AP). This led us to revisit the MQS method. The MQS approach selectively enhances positional queries with the top-k selected features while maintaining the learnability of content queries as before. We believe the primary reason for MQS's superior performance in baseline settings is its implicit incorporation of ranking information. This is an unintentional result of the specific code implementation of MQS. Specifically, by selecting the top-k initial positional queries, a sorting process is inherently involved. In the implementation, the highest-scoring positional query is always matched with the first learnable content query, and the lowest-scoring with the last, allow-

Method	Backbone	#epochs	AP	AP$_{50}$
Deformable++ [16]	R50	12	47.0	65.2
+MSelf-attention	R50	12	48.1	66.6
+MCross-attention	R50	12	48.3	67.0

Table 3. Component analysis. Our MSelf-attention and MCross-attention models are evaluated on a Deformable++ [16] baseline. Incorporating the MSelf-attention model results in an improvement of 1.1 AP. Further addition of the MCross-attention module introduce an additional increase of 0.2 AP.

Method	Backbone	#epochs	AP	AP$_{50}$
Deformable++ [16]	R50	12	47.0	65.2
w/R_{rank}	R50	12	47.8 (+0.8)	66.3
w/R_{rank}*R_{spt}	R50	12	48.1 (+1.1)	66.6

Table 4. Comparison of different self-attention adjustment strategies for EASE-DETR. Incorporating R_{rank} straightforwardly into our self-attention mechanism yielded an improvement of 0.8 AP. To further enhance locality, we multiply R_{rank} and R_{spt} to modulate self-attention, achieving an additional performance boost.

DP0	MQS	LFT	MSELF	AP
				43.7
			✓	46.9 (+3.2)
✓				44.3
✓			✓	47.4 (+3.1)
✓	✓			46.3
✓	✓		✓	47.7 (+1.4)
✓	✓	✓		47.0
✓	✓	✓	✓	48.1 (+1.1)

Table 5. Analysis of MSelf-Attention combined with various techniques. DP0: implementing a 0 Dropout Rate within the Transformer. MQS: mixed query selection. LFT: look forward twice. MSELF: MSelf-attention.

Method	#queries	GFLOPs	Params
DAB-DETR [21]	300	90.4	43.67 M
W/EASE	300	90.5 (+0.1)	43.67 M
DAB-DETR [21]	900	103.3	43.67 M
W/EASE	900	104.0 (+0.7)	43.67 M

Table 6. Analysis of the efficiency. Utilizing DAB-DETR [21] as our baseline, our EASE model incurs only a minimal increase of 0.1 GFLOPs and 0.7 GFLOPs for 300 and 900 queries respectively, with virtually no increase in the number of parameters.

Method	w/R_{rank}	NDS	mAP
PETR [24]		42.6	37.8
PETR [24]	✓	44.2 (+1.6)	38.7 (+0.9)

Table 7. Results on multi-view 3D object detection. By simply combining with R_{rank}, we achieve improvements of 1.6 NDS and 0.9 mAP on PETR [24], demonstrating its versatility across different DETR-based frameworks.

ing the learnable content queries to implicitly embody a sequence order through learning iterations.

Efficiency Analysis. In Table 6, we present an analysis of the computational costs introduced by our Mself-attention and Mcross-attention. We conducted comparative experiments with 300 and 900 queries respectively. At 300 queries, Our model only adds 0.1 GFLOPs, introducing a negligible amount of parameters, which is notably minimal. At 900 queries, due to the proportional relationship between computation and the number of queries, there is a more significant increase compared to 300 queries. However, the total increase of approximately 0.7 GFLOPs remains within acceptable limits.

Results on multi-view 3D object detection. Our approach is highly adaptable to various DETR-based methods and is not confined to 2D detection tasks. To demonstrate its universality, we apply our method to multi-view 3D detection tasks, as shown in Table 7. The results show that EASE-DETR achieves non-trivial improvements, *i.e.*, +1.6 NDS and +0.9 mAP on the PETR [24] baseline.

5. Conclusion

In conclusion, the EASE-DETR approach introduced in this study significantly enhances the non-duplicate detection capability of DETRs. By framing the detection process as a competition among object queries, where only one query emerges as the final detection, our method effectively manages the inherent challenges. The strategy of identifying and biasing towards the leading query in each decoder layer proves to be a crucial element in this improvement. This is accomplished by protecting the leading query from mutual suppression in the self-attention layer and bolstering its ability to assimilate more object features in the cross-attention layer. Consequently, this approach not only simplifies the competition but also accelerates the winning process of the leading query. Our experimental results demonstrate that EASE-DETR offers consistent and substantial enhancements across various DETR models.

6. Acknowledgments

This research is supported in part by National Key R&D Program of China (2022ZD0115502), National Natural Science Foundation of China (NO. 62122010, U23B2010), Zhejiang Provincial Natural Science Foundation of China under Grant No. LDT23F02022F02, Key Research and Development Program of Zhejiang Province under Grant 2022C01082.

References

[1] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 3

[2] Zhi Cai, Songtao Liu, Guodong Wang, Zheng Ge, Xiangyu Zhang, and Di Huang. Align-detr: Improving detr with simple iou-aware bce loss. *arXiv preprint arXiv:2304.07527*, 2023. 5, 6

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 2, 3

[4] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13039–13048, 2021. 3

[5] Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6633–6642, 2023. 5

[6] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2988–2997, 2021. 1

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[8] Jiahui Fu, Chen Gao, Zitian Wang, Lirong Yang, Xiaofei Wang, Beipeng Mu, and Si Liu. Eliminating cross-modal conflicts in bev space for lidar-camera 3d object detection. *arXiv preprint arXiv:2403.07372*, 2024. 3

[9] Chen Gao, Jinyu Chen, Si Liu, Luting Wang, Qiong Zhang, and Qi Wu. Room-and-object aware knowledge reasoning for remote embodied referring expression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3064–3073, 2021. 3

[10] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3621–3630, 2021. 3

[11] Yulu Gao, Chonghao Sima, Shaoshuai Shi, Shangzhe Di, Si Liu, and Hongyang Li. Sparse dense fusion for 3d object detection. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10939–10946. IEEE, 2023. 3

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 3

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5, 6

[14] Zhengdong Hu, Yifan Sun, Jingdong Wang, and Yi Yang. DAC-DETR: Divide the attention layers and conquer. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 5, 6

[15] Zhengdong Hu, Yifan Sun, and Yi Yang. Suppressing the heterogeneity: A strong feature extractor for few-shot segmentation. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[16] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023. 2, 5, 6, 8

[17] Borui Jiang, Ruixuan Luo, Jiayuan Mao, Tete Xiao, and Yuning Jiang. Acquisition of localization confidence for accurate object detection. In *Proceedings of the European conference on computer vision (ECCV)*, pages 784–799, 2018. 3

[18] 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*, pages 13619–13627, 2022. 1, 2

[19] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2, 6

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[21] 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. In *International Conference on Learning Representations*, 2021. 1, 2, 3, 8

[22] Shilong Liu, Tianhe Ren, Jiayu Chen, Zhaoyang Zeng, Hao Zhang, Feng Li, Hongyang Li, Jun Huang, Hang Su, Jun Zhu, et al. Detection transformer with stable matching. 2023. 2, 5, 6

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. In *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 3

[24] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022. 6, 8

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2, 6

[26] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang.

Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021. 1, 3

[27] Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, pages 850–855. IEEE, 2006. 3

[28] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 3

[30] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3611–3620, 2021. 1

[31] Zongheng Tang, Yifan Sun, Si Liu, and Yi Yang. Detr with additional global aggregation for cross-domain weakly supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11422–11432, 2023. 3

[32] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[34] Tao Wang, Li Yuan, Yunpeng Chen, Jiashi Feng, and Shuicheng Yan. Pnp-detr: Towards efficient visual analysis with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4661–4670, 2021. 1

[35] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2567–2575, 2022. 1

[36] Yue Wu, Yinpeng Chen, Lu Yuan, Zicheng Liu, Lijuan Wang, Hongzhi Li, and Yun Fu. Rethinking classification and localization for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10186–10195, 2020. 3

[37] Zizheng Xun, Shangzhe Di, Yulu Gao, Zongheng Tang, Gang Wang, Si Liu, and Bo Li. Linker: Learning long short-term associations for robust visual tracking. *IEEE Transactions on Multimedia*, 2024. 3

[38] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating detr convergence via semantic-aligned matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 949–958, 2022. 1

[39] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 5, 6

[40] Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7329–7338, 2023. 3

[41] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points, 2019. 3

[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020. 1, 2, 3