Relation DETR: Exploring Explicit Position Relation Prior for Object Detection

Xiuquan Hou[1], Meiqin Liu[1,2,*], Senlin Zhang[2] Ping Wei[1], Badong Chen[1], and Xuguang Lan[1]

[1] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'an 710049, China
[2] College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China
xiuqhou@stu.xjtu.edu.cn,{liumeiqin,slzhang}@zju.edu.cn,
{pingwei,chenbd,xglan}@mail.xjtu.edu.cn

Abstract. This paper presents a general scheme for enhancing the convergence and performance of DETR (DEtection TRansformer). We investigate the slow convergence problem in transformers from a new perspective, suggesting that it arises from the self-attention that introduces no structural bias over inputs. To address this issue, we explore incorporating position relation prior as attention bias to augment object detection, following the verification of its statistical significance using a proposed quantitative macroscopic correlation (MC) metric. Our approach, termed Relation-DETR, introduces an encoder to construct position relation embeddings for progressive attention refinement, which further extends the traditional streaming pipeline of DETR into a contrastive relation pipeline to address the conflicts between non-duplicate predictions and positive supervision. Extensive experiments on both generic and task-specific datasets demonstrate the effectiveness of our approach. Under the same configurations, Relation-DETR achieves a significant improvement (+2.0% AP compared to DINO), state-of-the-art performance (51.7% AP for 1× and 52.1% AP for 2× settings), and a remarkably faster convergence speed (over 40% AP with **only 2** training epochs) than existing DETR detectors on COCO val2017. Moreover, the proposed relation encoder serves as a universal plug-in-and-play component, bringing clear improvements for theoretically any DETR-like methods. Furthermore, we introduce a class-agnostic detection dataset, SA-Det-100k. The experimental results on the dataset illustrate that the proposed explicit position relation achieves a clear improvement of 1.3% AP, highlighting its potential towards universal object detection. The code and dataset are available at https://github.com/xiuqhou/Relation-DETR.

Keywords: Detection transformer · Object detection · Relation network · Progressive attention refinement · Feature enhancement

* Corresponding author

1 Introduction

Object detection aims to tackle the problems of bounding box regression and object classification for each object of interest. Recently, DEtection TRansformer (DETR) [4] has overcome the reliance on handcrafted designs of convolution detectors, achieving an elegant architecture in an end-to-end manner.

Despite exhibiting impressive detection performance on large-scale datasets such as COCO [35], their performances are prone to be influenced by dataset scale and suffer from slow convergence. The root cause of the problem is the conflict between non-duplicate predictions and positive supervision [23]. During the training process, DETR employs the Hungarian algorithm to assign a single positive prediction to each ground-truth for producing unique results. However, this leads to negative predictions dominating the majority of the loss function, causing insufficient positive supervision. Therefore, more samples and iterations are required for convergence. Previous attempts have explored the issue by introducing train-only architectures (*e.g.* query denoising [28], multiple groups of queries [6], auxiliary queries [23], collaborative hybrid assignment training [52]) for additional supervision or by incorporating hard mining in loss functions (*e.g.* IA-BCE loss [3], position-supervised loss [37]). Other works have proposed specific structures for better interaction between queries and feature maps (*e.g.* dynamic anchor query [36], cascade window attention [46]), as well as techniques to focus on high-quality queries (*e.g.* hierarchical filtering [20], dense distinct process [49] and query rank layer [40]). Despite these advancements, there has been little exploration of the issue from the perspective of self-attention, which is widely used in the transformer decoders in most DETR detectors.

The effectiveness of self-attention lies in its establishment of a high-dimensional relation representation among sequence embeddings [4, 42], which also serves as a key component for modeling relations among different detection feature representations [4]. However, such relation is an implicit representation since it assumes no structural bias over inputs, in which even position information is also needed to be learned from training data [33]. Consequently, the learning process of transformer is data-intensive and slow to converge. This analysis motivates us to introduce task-specific bias for realizing faster convergence and reducing data dependence.

In this paper, we explore enhancing DETR detectors from a novel perspective, namely explicit position relation prior. We first establish a metric for quantifying position relations in images, and analyze the distribution to verify its statistical significance. In this context, we introduce a position relation encoder to model all pairwise interactions between two bounding boxes, employing progressive attention refinement for cross-layer information interaction. To maintain the end-to-end property while providing sufficient positive supervision, we introduce a contrast relation strategy, which leverages both one-to-one and one-to-many matching while emphasizing the influence of position relation on deduplication. The proposed method is named as Relation-DETR.

Compared to previous works, the main feature of Relation-DETR is the integration of explicit position relation. In contrast, prior works focus on implicitly

learned attention weights from training data, leading to slow convergence. Intuitively, our proposed position relation can be seen as a plug-in-and-play design beneficial for non-duplication predictions, since it establishes a representation of relative positions among pairs of bounding boxes(similar to IoU in NMS [16]).

We evaluate the performance of Relation-DETR on the most popular object detection dataset, COCO 2017 [35], as well as several task-specific datasets [8,43]. The experimental results demonstrate its superior performance, surpassing previous state-of-the-art DETR detectors with clear margins. More specifically, Relation-DETR exhibits a significantly fast convergence speed. Without bells and whistles, it becomes the the first DETR detector to achieve 40% AP on COCO with only 2 epochs using ResNet50 as the backbone under the $1\times$ training configuration. In addition, the simple architecture design of our position relation encoder maintains a promising transferability. It can be easily extended to other DETR-based methods with only a few modifications to achieve consistent performance improvements. This is in contrast to some existing DETR detectors whose performance is highly dependent from complex matching strategies [22] or detection heads [52] developed by convolution-based detectors.

2 Related Work

Transformer for Object Detection In practice, the majority of attempts to apply transformer to object detection involve constructing a parallelizable sequence, either in the feature extractor [30] or in the detection body [4]. Specifically, transformer-based feature extractors generate token sequences based on image patches [12, 30, 39], and extract multi-scale features through aggregating local features [30, 39] or pyramid postprocess [14, 30]. DEtection TRansformer (DETR) proposed by Carion *et al.* [4] encodes the extracted features into object queries and decodes them into detected bounding boxes and labels. However, the self-learned attention mechanism increases the requirements for large-scale datasets and training iterations. Many works have explored the slow convergence from the perspective of structured attention (*e.g.* multi-scale deformable attention [51], dynamic attention [10], cascade window attention [46]), queries with explicit priors (*e.g.* anchor queries [44], dynamic anchor boxes queries [36], denoising queries [28], dense distinct queries [49]), and additional positive supervision (*e.g.* group queries [6], hybrid design [23], mixed matching [3]). However, even state-of-the-art DETR methods still utilize vanilla multi-head attention in the transformer decoder. And few works have explored the slow convergence from the perspective of implicit priors. This paper aims to address the issue with position relation.

Relation Network Rather than processing visual features at pixel levels, patch levels or image levels, relation network captures relation features at instance levels. Existing research on relation networks involves category-based and instance-based approaches. The category-based approaches construct conceptual or statistical relations (*e.g.* co-occurrence probability [17, 27]) either from relation

datasets like Visual Genome [24, 27, 45] or by adaptively learning from class labels [17]. Both of them, however, increase the complexity due to the assignment between instances and categories [7, 17, 45]. In contrast, instance-based approaches directly construct a fine-grained graph structure given object features as a node set and their relations as a edge set. Therefore, reasoning on the graph during the training process naturally determines the explicit relation weight [21]. Typically, the weight denotes the parametric distances between each paired object instances in high-dimensional space, such as appearance similarity [31], proposal distance [32] or even self-attention weight [42]. Since learning self-attention weight solely from training data without structural bias increases the requirement for dataset scale and iterations, we explore explicit position relation as a prior to reduce the requirement.

Classification loss for hard mining. During object detection training, positive predictions assigned to ground truth are much fewer than negative predictions, often resulting in imbalanced supervision and slow convergence. For classification tasks, Focal Loss [34] proposes introducing a weight parameter to focus on hard samples, which is further extended to many variants like generalized focal loss (GFL) [29], vari focal loss (VFL) [29]. Moreover, for object detection tasks, using loss with modulation terms based on regression metrics (*e.g.* TOOD [15], IA-BCE [3], position-supervised loss [37]) further achieves high-quality alignment between classification and regression tasks.

3 Statistical significance of object position relation

Are objects really correlated in object detection tasks? To answer the question, we propose a quantitative macroscopic correlation (MC) metric based on the Pearson Correlation Coefficient (PCC) to measure the position correlation among objects in a single image. Assume the objects in an image form a node set, and the PCC between each pair of bounding box annotations serves as their corresponding edge weight. We can construct an undirected graph with continuous values. In this end, the macroscopic correlation for each image can be calculated using the graph intensity, formulated as:

$$MC = \frac{\sum_i \sum_{j:j\neq i} |\texttt{Pearson}(\boldsymbol{b}_i, \boldsymbol{b}_j)|}{N(N-1)} \tag{1}$$

where N denotes the number of objects, *i.e.* the number of nodes, $\boldsymbol{b} = [x, y, w, h]$ denotes the position annotation of bounding boxes in datasets. $MC = 1$ only when all objects are fully linearly correlated, whereas $MC = 0$ if there is no position correlation between any pair of objects.

We visualize the statistical distribution of MC for datasets across various scenarios, including industrial (ESD [19], CSD [43], MSSD [8]), domestic (AI2thor [26]), urban (Cityscapes [9]) and generic settings (PascalVOC [13], COCO [35], Object365 [41], SA-1B [25]). The datasets cover a wide range of scales, from

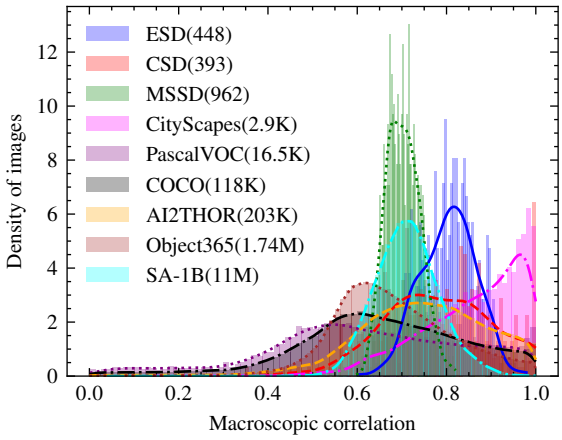

Fig. 1: Statistical distribution of macroscopic correlation (MC) on various datasets (normalized for better visualization), and the values in brackets indicate the number of dataset samples.

0.3K to 11M images. As shown in Fig. 1, all of these datasets indicate that the distribution of MC is concentrated around high numerical values, with the distribution centers closer to the upper bound. This demonstrates the presence and the statistical significance of object position relation. Specifically, task-specific datasets display more prior knowledge and clearer clustering patterns in the high-dimensional feature space, which thus results in higher MC values than generic datasets like COCO.

4 Relation-DETR

Given the statistical significance of position relation, we propose a state-of-the-art detector, named Relation-DETR, that explores explicit position relation prior to enhance object detection. To address the issue of slow convergence, we present a position relation encoder(Sec. 4.1) for progressive attention refinement(Sec. 4.2). Further in Sec. 4.3, it extends the streaming pipeline of DETR into a contrast pipeline to emphasize the influence of position relation on removing duplication while maintaining sufficient positive supervision for faster convergence.

4.1 Position relation encoder

Previous research has demonstrated the relation effectiveness for convolution detectors [21,32]. Recently, some DETR methods attempt to construct instance-level relation by indexing from category-level relation using class indices [17]. In contrast to these approaches, we directly construct instance-level relation through a simple position encoder, maintaining an end-to-end design for DETR.

We first review the basic pipeline in DETR detectors. Given image features extracted by the backbone, the transformer encoder produces an augmented memory $\mathbf{Z} \in \mathbb{R}^{d \times H \times W}$ for further decoding into bounding boxes $\boldsymbol{b}_i = [x, y, w, h], i = 1, \cdots, N$ and class labels $\boldsymbol{c}_i$ as predictions. Each decoder layer refines the bounding box coordinates iteratively by predicting Δ w.r.t. the coordinate from the last decoder layer, known as iterative bounding box refinement [51]. In addition, predictions from all decoder layers equally participate in loss calculation to compute auxiliary decoding losses [4].

Under the aforementioned detection framework, our position relation encoder represents the high-dimensional relation embedding as an explicit prior for self-attention in the transformer. This embedding is calculated based on the predicted bounding boxes (denoted as $\mathbf{b} = [x, y, w, h]$) from each decoder layer. To ensure that the relation is invariant to translation and scale transformations, we encode it based on normalized relative geometry features:

$$\boldsymbol{e}(\boldsymbol{b}_i, \boldsymbol{b}_j) = \left[\log\left(\frac{|x_i - x_j|}{w_i} + 1\right), \log\left(\frac{|y_i - y_j|}{h_i} + 1\right), \log\left(\frac{w_i}{w_j}\right), \log\left(\frac{h_i}{h_j}\right)\right] \quad (2)$$

Unlike [21], our position relation is unbiased, as $\boldsymbol{e}(\boldsymbol{b}_i, \boldsymbol{b}_j) = 0$ when $i = j$. The relation matrix $\mathbf{E} \in \mathbb{R}^{N \times N \times 4}$ (with $\mathbf{E}(i, j) = \boldsymbol{e}(\boldsymbol{b}_i, \boldsymbol{b}_j)$) is further transformed into high-dimensional embeddings through sine-cosine encoding [42].

$$\texttt{Embed}(\mathbf{E}, 2k) = \sin\left(s\mathbf{E}/T^{2k/d_{\text{re}}}\right) \quad (3)$$

$$\texttt{Embed}(\mathbf{E}, 2k+1) = \cos\left(s\mathbf{E}/T^{2k/d_{\text{re}}}\right) \quad (4)$$

where the shape of relation embedding is $N \times N \times 4d_{\text{re}}$, and T, d_{re}, s are encoding parameters. Finally, the embedding undergoes a linear transformation to obtain M scalar weights, where M denotes the number of attention heads.

$$\texttt{Rel}(\boldsymbol{b}, \boldsymbol{b}) = \max\left(\epsilon, \mathbf{W}\texttt{Embed}(\boldsymbol{b}, \boldsymbol{b}) + \mathbf{B}\right) \quad (5)$$

where ϵ makes sure a positive value for relation to avoid gradient vanishing after $\texttt{exp}$ when integrated into self-attention, and $\texttt{Rel}(\boldsymbol{b}, \boldsymbol{b}) \in \mathbb{R}^{N \times N \times M}$.

4.2 Progressive attention refinement with position relation

The iterative box refinement proposed by Deformable-DETR [51] has shown the effectiveness for high-quality bounding box regression. Following the motivation, we propose a progressive attention refinement method to introduce the position relation into the streaming pipeline of DETR. Specifically, the relation of layer-i is determined by bounding boxes of both layer-$i-1$ and layer-i, which is further integrated into self-attention for producing the bounding boxes in layer-$(i+1)$.

$$\texttt{Attn}_{\texttt{Self}}(\mathbf{Q}^l) = \texttt{Softmax}\left(\texttt{Rel}(\boldsymbol{b}^{l-1}, \boldsymbol{b}^l) + \frac{\texttt{Que}(\mathbf{Q}^l)\texttt{Key}(\mathbf{Q}^l)^\top}{\sqrt{d_{model}}}\right)\texttt{Val}(\mathbf{Q}^l) \quad (6)$$

$$\mathbf{Q}^{l+1} = \texttt{FFN}\left(\mathbf{Q}^l + \texttt{Attn}_{\texttt{cross}}\left(\texttt{Attn}_{\texttt{self}}(\mathbf{Q}^l), \texttt{Key}(\mathbf{Z}), \texttt{Val}(\mathbf{Z})\right)\right) \quad (7)$$

$$\boldsymbol{b}^{l+1} = \texttt{MLP}(\mathbf{Q}^{l+1}), \boldsymbol{c}^{l+1} = \texttt{Linear}(\mathbf{Q}^{l+1}) \quad (8)$$

where $\mathbf{Q}^l$ denotes the queries in the l-th decoder layer in the DETR transformer, $\mathbf{Z}$ is the memory, *i.e.*, the enhanced image features from the transformer encoder.

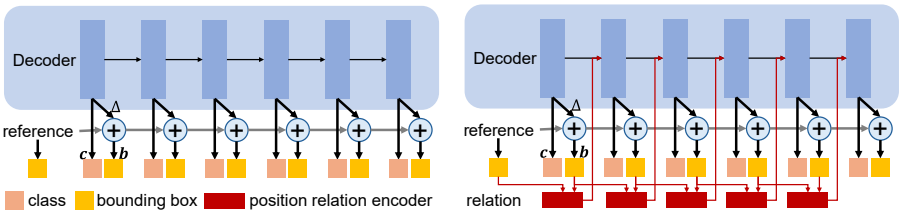

Fig. 2: Comparison of transformer decoder in Deformable-DETR(left) and Relation-DETR(right).

The only difference between our method and the existing DETR decoder is marked in red. As depicted in Fig. 2, the requisite addition involves the introduction of a lateral branch for the computation of position relation. Therefore, our position relation and the progressive attention refinement are straightforward, allowing for plug-in-and-play integration with self attention in existing DETR detectors to achieve consistent performance improvements (see Tab. 6).

4.3 Contrast relation pipeline

Rethinking the mechanism of existing duplication removal methods (including NMS [16], Soft-NMS [1], fast-NMS [2], Adaptive-NMS [38]), these processes heavily rely on IoU (intersection over Union), which, to some extent, signifies the position relation between bounding boxes. Therefore, we may hypothesize that integrating the position relation among queries in self-attention contributes to non-duplicated predictions in object detection, akin to [22].

The conflicts between non-duplicate predictions and sufficient positive supervision arise from the streaming pipeline of DETR, which must navigate between one-to-one matching and one-to-many matching. To overcome this limitation, we extend it to a contrast pipeline based on the proposed position relation. Specifically, we construct two parallel sets of queries, *i.e.* matching queries $\mathbf{Q}_m$ and hybrid queries $\mathbf{Q}_h$. Both are input into the transformer decoder but undergo distinct processing. The matching queries are processed with self-attention incorporating position relation to produce non-duplicated predictions:

$$\texttt{Attn}_{\texttt{Self}}(\mathbf{Q}^l_{\mathrm{m}}) = \texttt{Softmax}\left(\texttt{Rel}(\boldsymbol{b}^{l-1}, \boldsymbol{b}^l) + \frac{\texttt{Que}(\mathbf{Q}_{\mathrm{m}})\texttt{Key}(\mathbf{Q}_{\mathrm{m}})^\top}{\sqrt{d_{model}}}\right)\texttt{Val}(\mathbf{Q}_{\mathrm{m}}) \tag{9}$$

$$\texttt{Attn}_{\texttt{Self}}(\mathbf{Q}^l_{\mathrm{h}}) = \texttt{Softmax}\left(\frac{\texttt{Que}(\mathbf{Q}_{\mathrm{h}})\texttt{Key}(\mathbf{Q}_{\mathrm{h}})^\top}{\sqrt{d_{model}}}\right)\texttt{Val}(\mathbf{Q}_{\mathrm{h}}) \tag{10}$$

while the hybrid queries are decoded by the same decoder but skip the calculation of position relation to explore more potential candidates. Their corresponding predictions are denoted as $\boldsymbol{p}_m^l = (\boldsymbol{b}_m^l, \boldsymbol{c}_m^l)$ and $\boldsymbol{p}_h^l = (\boldsymbol{b}_h^l, \boldsymbol{c}_h^l)$, respectively. Details of the contrast relation pipeline are illustrated in Fig. 3.

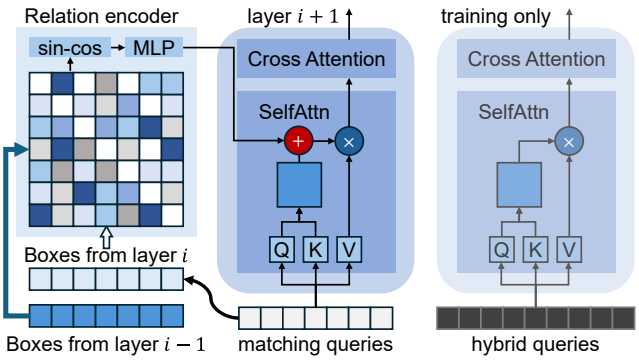

Fig. 3: Detailed illustration of the proposed contrast relation pipeline.

Assuming $\boldsymbol{g}$ denotes the ground truth annotations, for $\boldsymbol{p}_m$, we employ a one-to-one matching scheme to emphasize the non-duplicate property, with loss calculation similar to the original DETR approach [4]:

$$\mathcal{L}_m(\boldsymbol{p}_m, \boldsymbol{g}) = \sum_{l=1}^{L} \mathcal{L}_{\text{Hungarian}}(\boldsymbol{p}_m^l, \boldsymbol{g}) \tag{11}$$

While for $\boldsymbol{p}_h$, a one-to-many matching scheme is employed to form more potential positive candidates. We simply follow H-DETR [23] and repeat the ground truth K times, denoted as $\tilde{\boldsymbol{g}} = \{\boldsymbol{g}^1, \boldsymbol{g}^2, \cdots, \boldsymbol{g}^K\}$, for loss calculation:

$$\mathcal{L}_h(\boldsymbol{p}_h, \boldsymbol{g}) = \sum_{l=1}^{L} \mathcal{L}_{\text{Hungarian}}(\boldsymbol{p}_h^l, \tilde{\boldsymbol{g}}) \tag{12}$$

where $\mathcal{L}_{\text{Hungarian}}$ denotes the Hungarian loss, and L denotes the number of decoder layers. It is worth noting that the hybrid queries are only involved during training, thus incurring no extra computational burden for inference.

5 Experimental Results and Discussion

5.1 Setup

For a comprehensive evaluation, we conduct experiments on both an object detection benchmark, COCO 2017 [35], and two task-specific datasets, CSD [43]

and MSSD [8]. Detection performance is measured using the standard Average Precision (AP) [35]. We train our model on NVIDIA A800 GPU (80GB) and RTX 3090 GPU (24GB) using the AdamW optimizer with an initial learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. The Relation-DETR is implemented based on the backbone ResNet-50 [18] and Swin-L [39] pretrained from ImageNet [11], which are finetuned with a learning rate 1×10^{-5} during training. The learning rate is reduced by a factor of 0.1 at later stages. The parameters of the position relation encoder, T, d_{re}, s are empirically chosen as 10000, 16, 100, respectively. The hybrid training configurations follow DINO [47] and H-DETR [23], *i.e.*, $N_m = 900$, $N_h = 1500$, $k = 6$. We adopt the VariFocal loss [48] for training Relation-DETR. The training batch size is 10 for COCO 2017 and 2 for task-specific datasets. Before fed into detectors, images undergo the same augmentations (random resize, crop and flip) as other DETR detectors.

5.2 Comparison with state-of-the-art methods

Comparison on COCO 2017. Tab. 1 presents the detection performance on COCO `val` 2017. Compared to other state-of-the-art DETR methods, our approach converges much faster and demonstrates significant improvements of 1.0% AP, 1.0% AP_{50} and 0.6% AP_{75}, respectively, suppressing the second best DDQ-DETR [49] with clear margins. Specifically, Relation-DETR achieves 51.7% AP using only 12 epochs with the ResNet-50 backbone and even outperforms DINO [47] (51.2% AP) with 36 epochs (3× faster). More importantly, in contrast to DDQ-DETR [49] and $\mathcal{C}$o-DETR [52] that leverages NMS in the decoder or post-process for improving precision, our Relation-DETR maintains an end-to-end pipeline, ensuring promising extensibility. When integrated with Swin-L backbone, Relation-DETR outperforms all counterparts, achieving the best 57.8% AP with a 0.5% AP improvement, showcasing its excellent scalability for larger model capacity.

Comparison on task-specific datasets. Different from generic object detection benchmarks, datasets in task-specific scenarios lack sufficient samples to provide semantic information. To reveal the generalizability of Relation-DETR, we conduct a performance comparison on two defect detection datasets, *i.e.* CSD [43] and MSSD [8]. The results in Tab. 3 show that, Relation-DETR improves the baseline DINO by 1.4% AP on CSD, achieving the highest 54.4% AP. Tab. 4 demonstrates that Relation-DETR further increases the margin to 6.4% AP on MSSD and surpasses other counterparts. It is noteworthy that CSD [43] and MSSD [8] contain more small-sized objects than COCO 2017, which thus confirms the effectiveness of Relation-DETR to small-sized detection. Moreover, considering a stricter IoU threshold, Relation-DETR outperforms the second best method DINO by a significant margin of 11.1% AP@75, highlighting the beneficial impact of explicit position relation on high-quality predictions.

Table 1: Comparison with state-of-the-art methods on COCO `val2017` using ResNet-50(IN-1K) backbone. The $*$ means that we re-implement the methods and report the corresponding results.

Method	Backbone	Epochs	AP@50:95	AP@50	AP@75	AP_S	AP_M	AP_L
Anchor-DETR [44]	ResNet-50	50	42.1	63.1	44.9	22.3	46.2	60.0
Def-DETR* [51]	ResNet-50	12	45.4	65.0	49.1	27.2	49.6	61.0
DAB-Def-DETR* [36]	ResNet-50	12	46.3	65.9	50.4	29.3	50.5	61.9
DN-Def-DETR* [28]	ResNet-50	12	47.2	65.3	51.3	30.2	50.6	62.4
DINO* [47]	ResNet-50	12	49.9	67.4	54.5	33.9	53.5	63.8
Group-DETR [6]	ResNet-50	12	49.8			32.4	53.0	64.2
$\mathcal{H}$-Def-DETR [23]	ResNet-50	12	48.7	66.4	52.9	31.2	51.5	63.5
Cascade-DETR [46]	ResNet-50	12	49.7	67.1	54.1	32.4	53.5	65.1
$\mathcal{C}$o-Def-DETR [52]	ResNet-50	12	49.5	67.6	54.3	32.4	52.7	63.7
Align-DETR [3]	ResNet-50	12	50.2	67.8	54.4	32.9	53.3	65.0
Stable-DINO [37]	ResNet-50	12	50.4	67.4	55.0	32.9	54.0	65.5
DAC-DETR [22]	ResNet-50	12	50.0	67.6	54.7			
Salience-DETR* [20]	ResNet-50	12	50.0	67.7	54.2	33.3	54.4	64.4
Rank-DETR [40]	ResNet-50	12	50.4	67.9	55.2	33.6	53.8	64.2
MS-DETR [50]	ResNet-50	12	50.3	67.4	55.1	32.7	54.0	64.6
DDQ-DETR [49]	ResNet-50	12	50.7	68.1	55.7			
Relation-DETR	ResNet-50	12	**51.7**	**69.1**	**56.3**	**36.1**	**55.6**	**66.1**
DINO [47]	ResNet-50	24	50.4	68.3	54.8	33.3	53.7	64.8
DINO [47]	ResNet-50	36	51.2	69.0	55.8	35.0	54.3	65.3
$\mathcal{H}$-Def-DETR [23]	ResNet-50	36	50.0	68.3	54.4	32.9	52.7	65.3
Salience-DETR* [20]	ResNet-50	24	51.2	68.9	55.7	33.9	55.5	65.6
DDQ-DETR [49]	ResNet-50	24	52.0	69.5	**57.2**	35.2	54.9	65.9
Relation-DETR	ResNet-50	24	**52.1**	**69.7**	56.6	**36.1**	**56.0**	**66.5**

Table 2: Comparison with state-of-the-art methods on COCO `val2017` using Swin-L(IN-22K) as the backbone.

Method	Backbone	Epochs	AP@50:95	AP@50	AP@75	AP_S	AP_M	AP_L
DINO [47]	Swin-L	12	56.8	75.4	62.3	41.1	60.6	73.5
$\mathcal{H}$-Def-DETR [23]	Swin-L	12	55.9	75.2	61.0	39.1	59.9	72.2
Salience-DETR* [20]	Swin-L	12	56.5	75.0	61.5	40.2	61.2	72.8
Rank-DETR [40]	Swin-L	12	57.3	75.9	62.9	40.8	61.3	73.2
Rank-DINO [40]	Swin-L	12	57.6	76.0	**63.4**	**41.6**	61.4	73.8
Relation-DETR	Swin-L	12	**57.8**	**76.1**	62.9	41.2	**62.1**	**74.4**

5.3 Ablation study

This part conducts an ablation study to explore how the proposed components influence the final detection performance on COCO. The results in Tab. 5 show that, each key component of Relation-DETR consistently contributes to improving AP. Even on a highly-optimized baseline with VariFocal loss [48], our posi-

Table 3: Quantitative comparison on CSD [43].

Method	Backbone	Epochs	AP@50:95	AP@50	AP@75	AP_S	AP_M	AP_L
Def-DETR [51]	ResNet-50	300	43.7	86.2	36.6	40.5	34.9	**10.0**
DAB-Def-DETR [36]	ResNet-50	90	52.9	91.2	55.0	50.3	39.4	0.0
DN-Def-DETR [28]	ResNet-50	60	49.9	88.0	51.2	47.6	37.7	0.0
DINO [47]	ResNet-50	60	53.0	90.8	55.5	50.9	39.6	0.0
H-Def-DETR [23]	ResNet-50	60	53.0	90.6	55.7	51.2	39.2	6.7
Salience-DETR [20]	ResNet-50	60	53.2	92.5	55.1	51.0	40.9	0.0
Relation-DETR	ResNet-50	60	**54.4**	**92.9**	**56.0**	**53.6**	**40.8**	0.0

Table 4: Quantitative comparison on MSSD [8].

Method	Backbone	Epochs	AP@50:95	AP@50	AP@75	AP_S	AP_M	AP_L
Def-DETR [51]	ResNet-50	300	33.0	54.3	32.0	9.8	11.1	33.3
DAB-Def-DETR [36]	ResNet-50	120	33.7	60.0	31.0	15.9	26.7	29.0
DN-Def-DETR [28]	ResNet-50	120	45.6	74.1	44.9	18.6	31.9	41.0
H-Def-DETR [23]	ResNet-50	120	46.9	76.8	47.1	20.3	45.3	40.7
DINO [47]	ResNet-50	120	51.0	80.0	52.5	20.0	47.4	44.8
Salience-DETR [20]	ResNet-50	120	55.4	78.2	61.9	28.7	47.5	44.5
Relation-DETR	ResNet-50	120	**57.4**	**79.2**	**63.6**	**31.4**	**53.5**	**45.5**

tion relation encoder and contrast pipeline bring clear improvements of +0.3%, +0.5% AP, respectively. Built upon normalized relative geometry features, the position relation overcomes scale bias effectively, thus benefiting consistent performance improvements for different sized objects. For instance, Tab. 5 shows that introducing relation into the baseline with VFL achieves +1.2% AP_S, +1.0% AP_M, and +1.3% AP_L.

Table 5: Ablation study on key components of Relation-DETR (ResNet-50, 1x). We study both baseline and an optimized version with VariFocal Loss [48].

VFL	Rel.	contrast	AP@50:95	AP@50	AP@75	AP_S	AP_M	AP_L
			49.9	67.4	54.5	33.9	53.5	63.8
	✓		50.3(↑**0.4**)	68.1(↑**0.7**)	55.0(↑**0.5**)	34.7(↑**0.8**)	53.4(↓**0.1**)	64.7(↑**0.9**)
✓			50.9(↑**0.6**)	68.2(↑**0.1**)	55.3(↑**0.3**)	34.9(↑**0.2**)	54.6(↑**1.2**)	64.8(↑**0.1**)
✓	✓		51.2(↑**0.3**)	68.5(↑**0.7**)	55.7(↑**0.4**)	35.3(↑**0.4**)	54.9(↑**0.3**)	65.5(↑**0.7**)
✓	✓	✓	51.7(↑**0.5**)	69.1(↑**0.6**)	56.3(↑**0.6**)	36.1(↑**0.8**)	55.6(↑**0.7**)	66.1(↑**0.6**)

5.4 Transferability of position relation

Our position relation encoder adopts an elegant architectural design, ensuring a promising transferability to existing DETR detectors with minimal modifications. The experimental results in Tab. 6 show that, integrating position relation encoders without any further modifications enhances detection performance with clear margins of 1.6%, 2.0%, 0.1% and 0.2% for Deformable-DETR [51], DAB-Deformable-DETR [36], DN-Deformable-DETR [28] and DINO [47], respectively. Interestingly, in comparison to AP_M and AP_L, the position relation has a more substantial impact on improving AP_S for Deformable-DETR [51] and DAB-Deformable-DETR [36]. We attribute this to the fact that these early proposed baselines introduce relatively less structural bias and thus benefit more from our explicit position relation prior.

Table 6: Transfer experiments for the position relation encoder (ResNet-50, 1×). "+RelEnc" denotes the version that integrates our position relation encoder.

Method	AP@50:95	AP@50	AP@75	AP_S	AP_M	AP_L
Def-DETR	45.4	65.0	49.1	27.2	49.6	61.0
+RelEnc	47.0(↑**1.6**)	65.6(↑**0.6**)	51.2(↑**2.1**)	29.3(↑**2.1**)	51.0(↑**1.4**)	62.2(↑**1.2**)
DAB-Def-DETR	46.3	65.9	50.4	28.3	50.5	61.9
+RelEnc	48.3(↑**2.0**)	66.5(↑**0.6**)	52.9(↑**2.5**)	32.4(↑**4.1**)	52.0(↑**1.5**)	62.0(↑**0.1**)
DN-Def-DETR	47.2	65.3	51.3	30.2	50.6	62.4
+RelEnc	47.3(↑**0.1**)	65.6(↑**0.3**)	51.4(↑**0.1**)	29.9(↓**0.3**)	50.8(↑**0.2**)	62.1(↓**0.3**)

Moreover, the proposed contrast pipeline can be seen as an extension of hybrid matching [23], utilizing the proposed position relation encoder. Table 7 compares their transferability when integrated into DINO. The results indicate that directly applying hybrid matching [23] to DINO [47] results in a decrease in performance, from 49.9% AP to 49.5% AP. In contrast, the introduction of both the proposed relation encoder and the extended contrast pipeline consistently increases the performance. This demonstrates the effectiveness of the proposed position relation prior in improve detection performance, overcoming the weak generalizability inherent in hybrid matching.

Table 7: Transfer experiments compared with hybrid matching [23] on DINO.

DINO			+Hybrid [23]			+RelEnc			+RelEnc+Contrast		
AP	AP@50	AP@75	AP	AP@50	AP@75	AP	AP@50	AP@75	AP	AP@50	AP@75
49.9	67.4	54.5	49.5[†]	66.6	54.0	50.3	68.1	55.0	51.2	68.7	55.6

[†] We found that hybrid matching decreases the performance of DINO. The conclusion is consistent with the results of HDINO (see here) reported by MMDetection [5].

5.5 Intuitive performance comparison

To facilitate an intuitive performance comparison, Fig. 4 plots the convergence curve and the precision-recall curve. Because the position relation prior reduces the requirement for learning structural bias from data [33], Relation-DETR illustrates a faster convergence speed. When training from scratch, it achieves a higher AP than other counterparts with fewer iterations. Specifically, Relation-DETR can achieve over 40% AP with only **2** epochs, surpassing existing DETR detectors. In addition to convergence speed, the PR curves under different IoU thresholds also validate the performance gain of our Relation-DETR.

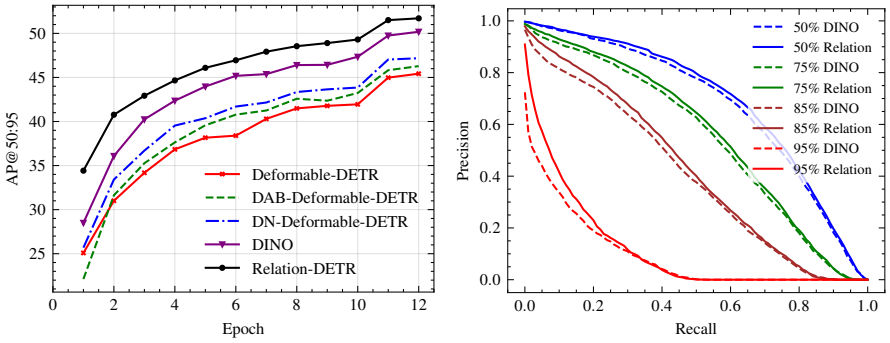

Fig. 4: Convergence curve(left) and Precision-recall curve for IoU=50% $\sim$ 95%(right). All models are trained with ResNet-50 backbone under the same 1$\times$ training configuration on COCO 2017.

5.6 Visualization

For a more intuitive grasp of the relation mechanism, Fig. 5 illustrates representative objects with high relation weights when given a query object. The visualization shows that for both generic and task-specific datasets, the relation contributes to identifying other detection candidates based on the given object query. Furthermore, small-sized objects tend to establish more relation connections with other objects due to the lack of their own semantic information. Therefore, constructing relations is crucial for small-sized object detection. Fig. 6 further visualizes some failure cases of Relation-DETR, indicating that the presented model may benefit from occluded objects and dense objects with misleading semantic differences by considering more complex relations, such as occlusion and semantic relations.

5.7 Towards universal object detection

Will the position relation prior still be effective for datasets that cover a broader range of scenarios and objects? As a piece of general prior knowledge, we an-

Fig. 5: Representative objects(red) related to the given object(blue)

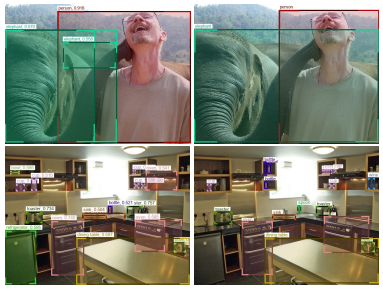

Fig. 6: Predictions (left) and GTs (right) of failure cases

ticipate that the explicit position relation prior can benefit universal object detection tasks. To explore this, we have constructed a large-scale class-agnostic detection dataset with about 100,000 images, termed SA-Det-100k, by sampling a subset from SA-1B, which is one of the largest scale segmentation datasets proposed in Segment Anything [25]. We then compared the performance of our Relation-DETR with the baseline DINO [47] using VFL [48] on this dataset. The results in Tab. 8 show that Relation-DETR achieves a clear improvement of 1.3% AP, demonstrating the scalability of the proposed position relation prior.

Table 8: Quantitative comparison on SA-Det-100k (ResNet-50, 1×).

Methods	AP@50:95	AP@50	AP@75	AP_S	AP_M	AP_L
DINO with VFL	43.7	52.0	47.7	5.8	43.0	61.5
Relation-DETR	45.0 **(↑1.3)**	53.1 **(↑1.1)**	48.9 **(↑1.2)**	6.0 **(↑0.2)**	44.4 **(↑1.4)**	62.9 **(↑1.4)**

6 Conclusion

This paper explores explicit position relation prior for enhancing performance and convergence of DETR detectors. Built upon normalized relative geometry features, we propose a novel position relation that overcomes scale bias for progressive attention refinement. To address the conflicts between non-duplicate predictions and sufficient positive supervision in DETR frameworks, we extend the streaming pipeline to a contrast pipeline based on the proposed position relation. Combining these components produces a state-of-the-art detector, named Relation-DETR. Massive ablation studies and experimental results demonstrate the superior performance, faster convergence and promising transferability of the proposed detector. Moreover, Relation-DETR exhibits remarkable generalizability for both generic and task-specific detection tasks. We believe that this work will inspire future research on relation and structural bias for DETR detectors.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 62327808, and the Fundamental Research Funds for Xi'an Jiaotong University under Grants xtr072022001 and xzy022024009.

References

1. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS–improving object detection with one line of code. In: Int. Conf. Comput. Vis. pp. 5561–5569 (2017)
2. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT: Real-time instance segmentation. In: Int. Conf. Comput. Vis. pp. 9157–9166 (2019)
3. Cai, Z., Liu, S., Wang, G., Ge, Z., Zhang, X., Huang, D.: Align-DETR: Improving DETR with simple IoU-aware BCE loss. arXiv preprint arXiv:2304.07527 (2023)
4. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Eur. Conf. Comput. Vis. pp. 213–229. Springer (2020)
5. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., et al.: Mmdetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019)
6. Chen, Q., Chen, X., Wang, J., Zhang, S., Yao, K., Feng, H., Han, J., Ding, E., Zeng, G., Wang, J.: Group DETR: Fast DETR training with group-wise one-to-many assignment. In: Int. Conf. Comput. Vis. pp. 6633–6642 (2023)
7. Chen, X., Li, L.J., Fei-Fei, L., Gupta, A.: Iterative visual reasoning beyond convolutions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7239–7248 (2018)
8. Chen, Y., Pan, J., Lei, J., Zeng, D., Wu, Z., Chen, C.: EEE-Net: Efficient edge enhanced network for surface defect detection of glass. IEEE Transactions on Instrumentation and Measurement (2023)
9. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3213–3223 (2016)
10. Dai, X., Chen, Y., Yang, J., Zhang, P., Yuan, L., Zhang, L.: Dynamic DETR: End-to-end object detection with dynamic attention. In: Int. Conf. Comput. Vis. pp. 2988–2997 (2021)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 248–255. Ieee (2009)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2020)
13. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**, 303–338 (2010)
14. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: EVA-02: A visual representation for neon genesis. arXiv preprint arXiv:2303.11331 (2023)
15. Feng, C., Zhong, Y., Gao, Y., Scott, M.R., Huang, W.: TOOD: Task-aligned one-stage object detection. In: Int. Conf. Comput. Vis. pp. 3490–3499. IEEE Computer Society (2021)
16. Girshick, R.: Fast R-CNN. In: Int. Conf. Comput. Vis. pp. 1440–1448 (2015)

17. Hao, X., Huang, D., Lin, J., Lin, C.Y.: Relation-enhanced DETR for component detection in graphic design reverse engineering. In: IJCAI. pp. 4785–4793 (2023)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 770–778 (2016)
19. Hou, X., Liu, M., Zhang, S., Wei, P., Chen, B.: CANet: Contextual information and spatial attention based network for detecting small defects in manufacturing industry. Pattern Recognition **140**, 109558 (2023)
20. Hou, X., Liu, M., Zhang, S., Wei, P., Chen, B.: Salience detr: Enhancing detection transformer with hierarchical salience filtering refinement. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17574–17583 (2024)
21. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3588–3597 (2018)
22. Hu, Z., Sun, Y., Wang, J., Yang, Y.: DAC-DETR: Divide the attention layers and conquer. Adv. Neural Inform. Process. Syst. **36** (2024)
23. Jia, D., Yuan, Y., He, H., Wu, X., Yu, H., Lin, W., Sun, L., Zhang, C., Hu, H.: DETRs with hybrid matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19702–19712 (2023)
24. Jiang, C., Xu, H., Liang, X., Lin, L.: Hybrid knowledge routed modules for large-scale object detection. Adv. Neural Inform. Process. Syst. **31** (2018)
25. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4015–4026 (2023)
26. Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., Deitke, M., Ehsani, K., Gordon, D., Zhu, Y., et al.: Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474 (2017)
27. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis. **123**, 32–73 (2017)
28. Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L.: DN-DETR: Accelerate DETR training by introducing query denoising. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13619–13627 (2022)
29. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. Adv. Neural Inform. Process. Syst. **33**, 21002–21012 (2020)
30. Li, Y., Mao, H., Girshick, R., He, K.: Exploring plain vision transformer backbones for object detection. In: Eur. Conf. Comput. Vis. pp. 280–296. Springer (2022)
31. Li, Z., Du, X., Cao, Y.: Gar: Graph assisted reasoning for object detection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1295–1304 (2020)
32. Lin, J., Pan, Y., Lai, R., Yang, X., Chao, H., Yao, T.: Core-Text: Improving scene text detection with contrastive relational reasoning. In: Int. Conf. Multimedia and Expo. pp. 1–6. IEEE (2021)
33. Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. AI Open (2022)
34. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Int. Conf. Comput. Vis. pp. 2980–2988 (2017)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Eur. Conf. Comput. Vis. pp. 740–755. Springer (2014)

36. Liu, S., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., Zhu, J., Zhang, L.: DAB-DETR: Dynamic anchor boxes are better queries for DETR. In: Int. Conf. Learn. Represent. (2021)
37. Liu, S., Ren, T., Chen, J., Zeng, Z., Zhang, H., Li, F., Li, H., Huang, J., Su, H., Zhu, J., et al.: Detection transformer with stable matching. arXiv preprint arXiv:2304.04742 (2023)
38. Liu, S., Huang, D., Wang, Y.: Adaptive NNS: Refining pedestrian detection in a crowd. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6459–6468 (2019)
39. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Int. Conf. Comput. Vis. pp. 10012–10022 (2021)
40. Pu, Y., Liang, W., Hao, Y., Yuan, Y., Yang, Y., Zhang, C., Hu, H., Huang, G.: Rank-DETR for high quality object detection. Adv. Neural Inform. Process. Syst. **36** (2024)
41. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Int. Conf. Comput. Vis. pp. 8430–8439 (2019)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
43. Wang, Q., Gao, S., Xiong, L., Liang, A., Jiang, K., Zhang, W.: A casting surface dataset and benchmark for subtle and confusable defect detection in complex contexts. IEEE Sensors Journal pp. 1–1 (2024)
44. Wang, Y., Zhang, X., Yang, T., Sun, J.: Anchor DETR: Query design for transformer-based detector. In: AAAI. vol. 36, pp. 2567–2575 (2022)
45. Xu, H., Jiang, C., Liang, X., Lin, L., Li, Z.: Reasoning-RCNN: Unifying adaptive global reasoning into large-scale object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6419–6428 (2019)
46. Ye, M., Ke, L., Li, S., Tai, Y.W., Tang, C.K., Danelljan, M., Yu, F.: Cascade-DETR: Delving into high-quality universal object detection. In: Int. Conf. Comput. Vis. pp. 6704–6714 (2023)
47. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: Int. Conf. Learn. Represent. (2022)
48. Zhang, H., Wang, Y., Dayoub, F., Sunderhauf, N.: Varifocalnet: An iou-aware dense object detector. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8514–8523 (2021)
49. Zhang, S., Wang, X., Wang, J., Pang, J., Lyu, C., Zhang, W., Luo, P., Chen, K.: Dense distinct query for end-to-end object detection. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7329–7338 (2023)
50. Zhao, C., Sun, Y., Wang, W., Chen, Q., Ding, E., Yang, Y., Wang, J.: Ms-detr: Efficient detr training with mixed supervision. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 17027–17036 (2024)
51. Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J.: Deformable DETR: Deformable transformers for end-to-end object detection. In: Int. Conf. Learn. Represent. (2020)
52. Zong, Z., Song, G., Liu, Y.: DETRs with collaborative hybrid assignments training. In: Int. Conf. Comput. Vis. pp. 6748–6758 (2023)