Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(8): 14158–14179.
DOI: 10.3934/mbe.2023633
Received: 12 April 2023
Revised: 23 May 2023
Accepted: 15 June 2023
Published: 26 June 2023

Research article

AD-DETR: DETR with asymmetrical relation and decoupled attention in crowded scenes

Yueming Huang[1,2] **and Guowu Yuan**[1,2,*]

[1] School of Information Science and Engineering, Yunnan University, Kunming 650504, China

[2] Yunnan Key Laboratory of Intelligent Systems and Computing, Kunming 650504, China

* **Correspondence:** Email: gwyuan@ynu.edu.cn; Tel: +8687165033748.

Abstract: Pedestrian detection in crowded scenes is widely used in computer vision. However, it still has two difficulties: 1) eliminating repeated predictions (multiple predictions corresponding to the same object); 2) false detection and missing detection due to the high scene occlusion rate and the small visible area of detected pedestrians. This paper presents a detection framework based on DETR (detection transformer) to address the above problems, and the model is called AD-DETR (asymmetrical relation detection transformer). We find that the symmetry in a DETR framework causes synchronous prediction updates and duplicate predictions. Therefore, we propose an asymmetric relationship fusion mechanism and let each query asymmetrically fuse the relative relationships of surrounding predictions to learn to eliminate duplicate predictions. Then, we propose a decoupled cross-attention head that allows the model to learn to restrict the range of attention to focus more on visible regions and regions that contribute more to confidence. The method can reduce the noise information introduced by the occluded objects to reduce the false detection rate. Meanwhile, in our proposed asymmetric relations module, we establish a way to encode the relative relation between sets of attention points and improve the baseline. Without additional annotations, combined with the deformable-DETR with Res50 as the backbone, our method can achieve an average precision of 92.6%, MR^{-2} of 40.0% and Jaccard index of 84.4% on the challenging CrowdHuman dataset. Our method exceeds previous methods, such as Iter-E2EDet (progressive end-to-end object detection), MIP (one proposal, multiple predictions), etc. Experiments show that our method can significantly improve the performance of the query-based model for crowded scenes, and it is highly robust for the crowded scene.

Keywords: pedestrian detection; crowded object detection; DETR; end-to-end detector; crowded pedestrian scene; relation net; attention mechanism; symmetry

1. Introduction

1.1. Background

Object detection in crowded scenes is widely used in self-driving, surveillance and robotics. There are two challenges for object detection in crowded scenes. One is the elimination of dense repetitive predictions, and the other is high false detection due to high occlusion rates. A good detector must detect all objects while avoiding repetitive prediction.

To solve these problems, one-stage [1–3] and two-stage detectors [4–6] use dense-to-dense and dense-to-sparse structures to generate dense candidate boxes or sparse candidate boxes by setting dense anchors, respectively. The matching strategy achieves higher recall by assigning multiple candidates to one ground truth (GT). Since the network itself cannot de-duplicate, the higher recall will result in many repeated predictions, so these methods require additional post-processing to de-duplicate. De-duplication refers to the following: when multiple predictions correspond to the same GT, the model can eliminate the other repeated predictions while reserving the best one. Those anchor-based approaches use non-maximum suppression (NMS) [6,7] based on greedy algorithms or improved post-processing methods as the de-duplication module. In the NMS algorithm, different intersection over union (IOU) threshold configurations is a trade-off strategy. A high threshold increases accuracy but decreases recall and reduces repeated predictions; a low threshold raises the recall rate but decreases accuracy and generates a large number of repeated predictions. At the same time, when the IOU between GTs is higher than the IOU threshold specified by NMS, there will inevitably be missed detection. AdaptiveNMS [6] introduces an additional network to learn the density of GTs in different scenes to dynamically modify the NMS threshold in real time to adapt to scenes with different densities. Soft-NMS [7] improves the original NMS, directly removes the GT with high surrounding overlap and reduces the surrounding prediction confidence, but it is still a greedy algorithm. In the dense object scene, the detectors based on greedy algorithm post-processing still have serious repeated prediction and missed detection.

A series of end-to-end query-based detection methods [8–10], represented by DETR [11] and sparse-RCNN (sparse region convolutional neural network) [12], regard target detection as a set prediction problem and integrate the de-duplication ability into the network through the end-to-end learning prediction of bipartite graph matching. These methods avoid the repeated prediction and missed detection caused by the manual anchor setting and NMS based on the greedy algorithm. The improved methods of DETR [8–10, 13] solve the slow convergence, low accuracy and poor interpretability of the basic model of DETR. These improved methods have achieved state-of-the-art results on some famous general datasets such as COCO [14], PASCAL VOC [15], etc. They have significantly improved the performance compared with the previous framework [1, 16] of NMS post-processing methods based on greedy algorithms.

However, these DETR-like models generally use a symmetrical decoder structure, which means that queries can equally pay attention to and gather information from each other in networks. As shown in Figure 1, we found that in a symmetrical structure, the prediction boxes at each stage tend to adopt the same strategy to update themselves, because they can sense each other equally. In the left of Figure 1, both prediction boxes *a* and *b* reduced their confidence and gave up regression to the GT, resulting in missed detection; In the right of Figure 1, both prediction boxes increased their confidence to regress to the GT, resulting in repeated predictions; we call it the synchronous update problem.

Figure 1. The synchronous update of the prediction boxes in the fine-tuning stage. The solid boxes *a* and *b* represent the prediction boxes in the current fine-tuning stage by the decoder blocks, and the dashed boxes represent the possible synchronous updates after fine-tuning by the symmetric decoder structure.

Since the synchronous update problem is caused by a symmetrical decoder structure in the fine-tuning stage, the de-duplication ability learned by the network is still limited. Moreover, DETR-like models generally use a fixed attention range. Because of the high occlusion rate of GTs in crowded scenes, such a strategy will introduce a lot of noise from the occlusion part to the process of predicting. The visualization of the attention points of deformable DETR is shown in Figure 2; note that, due to the high occlusion between GTs in a crowded scene, only a few attention points are located in the visible area of the GT. In contrast, most attention points fall on the surrounding occluded objects, introducing a large amount of noise into confidence prediction. The detection results of DETR and its improvements on the highly crowded dataset *CrowdHuman* [17] show that the de-duplication ability learned by the network itself and its adaptability to crowded scenes are still limited.

In those optimization works on query-based detectors in crowded scenes, Zhou and Yuan [18] eliminated repeated predictions by allowing the model to regress the full-body and visible-region boxes using additional visible-region annotations. A series of methods based on this idea used the visible prediction boxes with a low repetition rate to assist the full-body box predictions with a high repetition rate in crowded scenes [19–21]. The pedestrian end-to-end detector (PED) [22] reduced missing detection by guiding the cross-attention range by annotating the visible box region. However, these methods required additional costs for visible-box labeling and had limited improvement in performance. Iter-E2EDet [23] regarded prediction box de-duplication as a de-noising task. It considered the predictions higher than the fixed confidence threshold as accepted predictions and other predictions with low confidence as noise to eliminate duplication. However, this method was effective only when repeated predictions are between low-confidence and high-confidence predictions. This method had no effect when repeated predictions were between low-confidence predictions or high-confidence predictions.

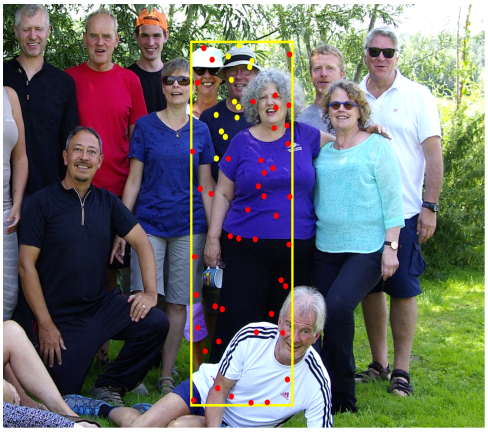


Figure 2. Confidence prediction of occluded GT (yellow attention points are attention points that contribute to confidence prediction in the visible area, and red attention points are noise information).

1.2. Our motivations

The DETR-like detection framework can learn end-to-end de-duplication and has competitive state-of-the-art results on the standard datasets [14, 15]. Therefore, we follow the overall design of deformable DETR [9], but we divide the decoders into decoupled decoder blocks and asymmetrical relation decoder blocks.

In decoupled decoders, we can reduce missed and false predictions by narrowing the attention range for confidence prediction to eliminate the noise from the occlusion decoupled decoder. Asymmetrical relation decoder blocks are designed to break the model's symmetry, which can eliminate duplicate predictions by solving simultaneous updates in the fine-tuning stage.

Figure 3 shows our overall framework. The feature extractor includes backbone and transformer encoders, and the cascaded decoders query the extracted feature maps to make predictions. In the decoupled encoder blocks, we decouple the information of confidence prediction and regression tasks. The regression output of the last decoupled encoder is a prediction of the position of the detection boxes, and the output for confidence prediction is fed into the following asymmetrical block. In the asymmetrical block, we propose an asymmetrical relation net(ARN) to break the model's symmetry, which can avoid the synchronous update of the prediction boxes to eliminate duplicate predictions.

Asymmetrical relation module. To solve the synchronous update problem caused by a symmetrical decoder structure, we adopted the idea of Iter-E2EDet [23], which embeds the relationship network in the model to integrate the relative relationship information between predictions. In our proposed asymmetrical relation decoder, we have introduced an asymmetric prediction information fusion mechanism based on confidence. Each prediction only pays attention to and gathers information on those predictions with a higher confidence score than it. Thus, this mechanism breaks the symmetry of the original model and avoids repeated or missed prediction caused by the synchronous update.

Decoupled confidence prediction. In the widely used DETR framework, a lot of noise information from the occlusion part will be introduced by a fixed attention range design of the decoder block. Moreover, the same attention point set information is used for both confidence and location predictions. However, the information required for the two types of prediction is different; confidence prediction

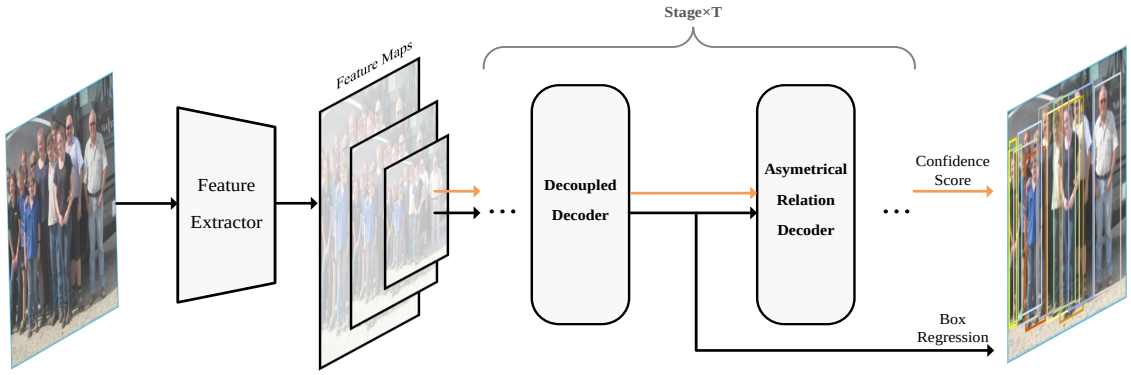


Figure 3. The overall framework of our method. We use the decoupled decoders to restrict the range of attention to let the model focus more on visible regions. At the end of our model, we use the asymmetrical relation decoders to break the model's symmetry to avoid the synchronous update of the prediction boxes in the fine-tuning stage.

tends to pay more attention to essential areas, such as the visible or head areas. In contrast, location prediction tends to focus on the location information of the prediction box boundary. Based on the above analysis, in our symetrical decoupled block, we propose a decoupled cross-attention head to separate the attention points of confidence and position predictions without any extra parameters added. The decoupled cross-attention head forces the model to only select a small number of attention points for confidence prediction. Therefore, our model can self-adaptively learn to focus on visible areas that contribute more to confidence prediction. By this design, our model can reduce the amount of noise information from occluded parts, thus reducing the missed detection rate.

Point set relation encoding. Due to the high spatial complexity of traditional sine and cosine spatial positional encoding [11], it is difficult to encode the relative relationship between the predicted attention point sets. Therefore, we propose a network to encode the relative relationship between point sets. Using this encoding method, we also achieved significant improvement on the baseline.

1.3. Our contribution

On the challenging CrowdHuman dataset [17], our experiments demonstrate that our proposed method can achieve state-of-the-art results. At the same time, we have analyzed the detection results in different density scenarios and prove the generalization performance in different density scenarios. The main contributions of this paper are as follows:

1) We propose an asymmetric relation mechanism to solve the synchronous update problem caused by the model's symmetry and reduce repeated predictions in crowded scenes.
2) We propose a simple and effective decoupled multi-head cross-attention algorithms to reduce noise information from surrounding occluded objects without adding any parameters, reducing the rate of missed and false predictions.

2. Review of DETR-like detectors

This section reviews the DETR-like detectors and their decoder structure for cascade refining. In the DETR-like models, the backbone first extracts the image features to form the image sequence information as the subsequent input. Second, the encoder blocks encode the sequence information with global attention to create multi-scale feature maps. Third, the decoder blocks update the position and confidence of the predictions by integrating the multi-scale feature maps in a cascading way. Finally, the classification head and regression head obtain the final prediction results. A decoder block in the DETR model can be formulated as follows:

$$\begin{aligned} q_t'' &\leftarrow MSA(q_{t-1}) + q_{t-1} \\ q'_t &\leftarrow MCA(q''_t, e_f) + q_t'' \\ q_t &\leftarrow ffn(q'_t) \\ box_t &\leftarrow \mathcal{B}(q_t) \\ cls_t &\leftarrow C(q_t) \end{aligned} \tag{2.1}$$

where $q \in \mathbb{R}_{N\times d}$ denotes the learnable object query; N and d denote the number and dimension of query q, respectively. The decoder block at the stage t receives the queries q_{t-1} output from the block at $t-1$ layer as an input. Muti-head self-attention $MSA(\cdot)$ is applied to perform self-attention among q_{t-1} to generate a multi-scale feature map e_f. Then, the multi-scale feature map information is integrated by the muti-head cross-attention $MCA(\cdot)$ and feed-forward networks $ffn(\cdot)$ to get the query q_t for the next stage. Simultaneously, q_t is fed into the box prediction branch $\mathcal{B}(\cdot)$ and classification branch $C(\cdot)$ for the current bounding box prediction box_t and confidence score prediction cls_t.

In the improved deformable DETR [9], the original global muti-head cross attention is replaced with deformable muti-head cross attention by adding the prior knowledge of the localization of the image. The improved deformable muti-head cross attention $MCA(q''_t, e_f|box_{t-1})$ can be formulated as follows:

$$q'_t = \sum_{n}^{N} W_n \left[\sum_{k=1}^{K} A_{nqk} \cdot W_n' e_f(p_q + \Delta p_{nqk}) \right] \tag{2.2}$$

where n indexes the attention head, $W_n \in C \times C_v$ ($C_v = C/N$) represent the learnable weights, the attention weight A_{nqk} is normalized by $\sum_{k=1}^{K} A_{nqk} = 1$, where $q \in \mathbb{R}_{N\times d}$ denotes the learnable object query and e_f is the multi-scale feature map encoded by encoder blocks. p_q is the coordinate of the center point of the bounding box corresponding to the query q, and Δp_{nqk} is the offset of the k^{th} attention point in the n^{th} attention head.

3. Our approach

Due to the excellent convergence and low computational complexity of deformable attention [9] and the ability to fuse multi-scale feature maps, we chose deformable DETR [9] as our implementation basis for our default instantiation. This section focuses on the detailed implementation of our method with a deformable DETR base, and our work can be applied to most query-based object detectors.

Figure 4 shows our approach's detailed implementation framework, including the asymmetrical relation decoders with ARN and decoupled decoders with decoupled multi-head cross attention

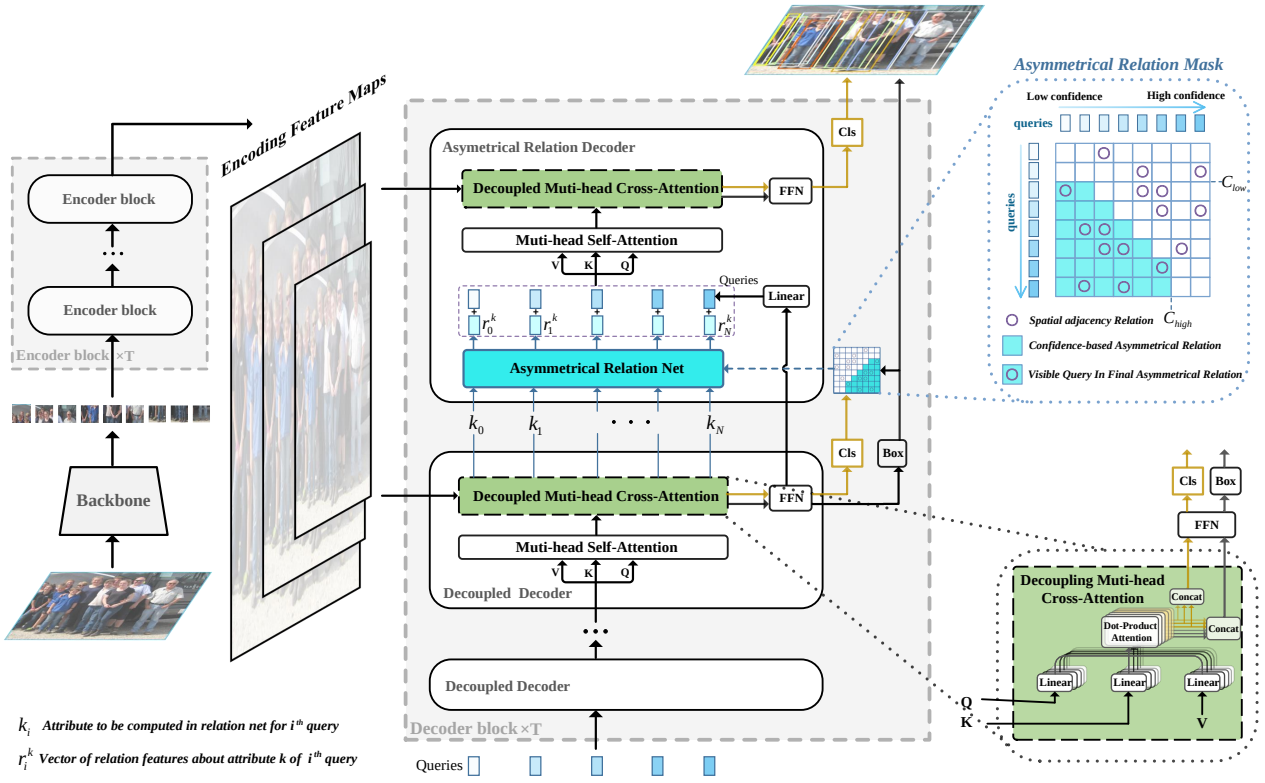


Figure 4. The diagram of our proposed DETR detection framework. It includes asymmetric decoders with ARN and decoupled decoders with *DMCA*.

(DMCA). The framework first feeds the image into the backbone. Then, the flattened output is transmitted into the transformer encoder to get multi-scale feature maps, which the decoder blocks will query to refine the predictions hierarchically.

In the decoupled decoders, we replace the original cross attention with our proposed DMCA to decouple information for confidence and regression prediction. In the final asymmetrical decoder, the asymmetric relation mask and the attribute k of each query from the last decoupled decoder are first fed into the ARN to obtain the asymmetrical relation vector r^k of each query. Asymmetrical relation vector r^k is a vector of the same dimension as the query to provide information on the surrounding predictions; the attribute k denotes the positions of boxes or attention points which are utilized by the ARN to form the relation vector. Then, those queries summed with the relation vector are fed into the DMCA and feed-forward network (FFN) later. Finally, the prediction results are obtained through the regression and classification heads. One improved asymmetrical decoder block can be formulated as follows:

$$
\begin{aligned}
r^k &\leftarrow \mathcal{AR}^k(P^k|M) \\
{q_t}^{rel} &\leftarrow fcn(r^k + fcn\,(q_{t-1})) \\
{q_t}'' &\leftarrow MSA({q_t}^{rel}) + {q_t}^{rel} \\
{q'}_t, {q'}_t^c &\leftarrow DMCA({q''}_t, p_{t-1}, e_{\mathrm{f}}) + {q_t}'' \\
q_t, q_t^c &\leftarrow ffn({q'}_t, {q'}_t^c) \\
box_t &\leftarrow \mathcal{B}(q_t) \\
\mathrm{cls_t} &\leftarrow C(q_t^c)
\end{aligned}
\tag{3.1}
$$

where $\mathcal{AR}^k(\cdot)$ denotes our proposed asymmetric relation module to encode the asymmetric relation of attribute k of the surrounding predictions, P^k represents the set of attributes k of queries and M represents the mask of asymmetric relation. $fcn(\cdot)$ denotes the fully connected layer. ${q_t}^{rel}$ represents the query after incorporating the asymmetric relation vector r^k. p_{t-1} is the centroid coordinates of the prediction box corresponding to the query at stage $t-1$. $DMCA(\cdot)$ is the DCMA module, whose outputs ${q'_t}^c$ and q'_t are fed into the FFN with the same parameters to obtain q_t and ${q_t}^c$, and finally the predictions box_t and cls_t of this stage are outputted through the regression head $\mathcal{B}$ and classification head C by ${q'_t}^c$ and q'_t, respectively.

3.1. Asymmetrical relation

As shown in Figure 1, the decoders' symmetry of the original deformable-DETR can bring about the synchronous update problem of queries, leading to duplicate predictions. We propose an asymmetric relational module $\mathcal{AR}^k$ to fuse the relation of attributes k among queries asymmetrically. That is, each prediction only focuses on and fuses the information of surrounding predictions with higher confidence than itself in a one-way manner. At the fine-tuning stage, $\mathcal{AR}^k$ can asymmetrically update the predictions to solve the synchronous update problem, thus eliminating duplicate predictions. The asymmetric relation module can be formulated as follows:

$$
r^k = \mathcal{AR}^k(P^k|M) \tag{3.2}
$$

The ARN accepts the set of attribute k of N queries $P^k = \{k_0, k_1, \ldots, k_N\}$ which denotes the positions of prediction boxes or attention points of queries and outputs asymmetric relation vector r^k.

We use the matrix $M = [m_{ij}]_{N\times N}$ to denote the asymmetric relation mask for N queries in the $\mathcal{AR}$ module, where $m_{ij} = 1$ means that q_i can notice q_j; $m_{ij} = 0$ means that q_i cannot notice q_j and no relation is computed. The relation matrix M can be decomposed into the Hadamard product of the spatial adjacency relation M^N and the confidence-based asymmetric relation M^A. So, we have that $M = M^N \circ M^A$, where $\circ$ means Hadamard product and M^N and M^A are computed as follows:

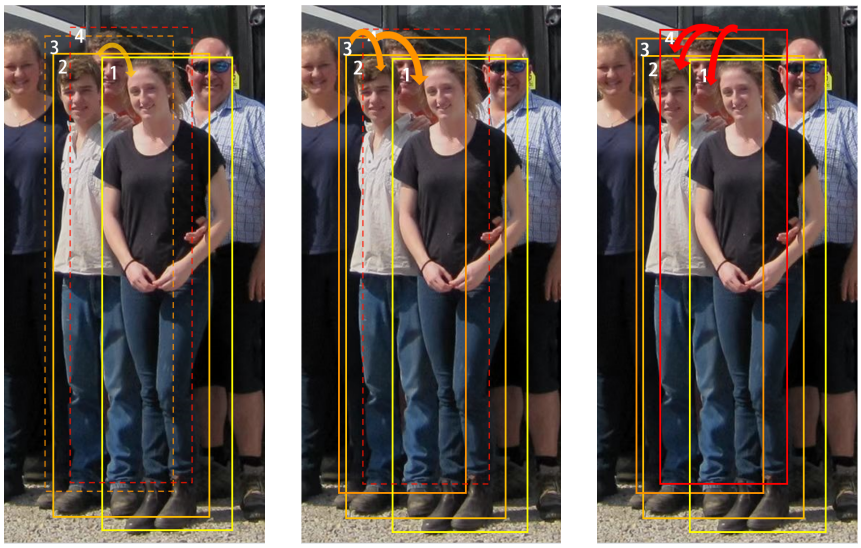
$$
m_{ij}^N = \begin{cases} 1, & IoU_{ij} > I_{thresh} \\ 0, & otherwise. \end{cases} \tag{3.3}
$$

$$
m_{ij}^A = \begin{cases} 1, & \text{if } c_i < c_j \text{ and } c_j > C_{low} \\ & \text{and } c_i < C_{high} \\ 0, & otherwise. \end{cases} \tag{3.4}
$$

where c_i represents the class confidence score of the i^{th} query. We set a minimum confidence threshold C_{low} to avoid incorporating much noise from low-confidence predictions. We also set the upper confidence limit C_{high} to reduce computation, because we found that the accuracy of high-confidence predictions is often already very high. In the confidence-based asymmetric relationship M^A, each prediction can only one-way notice the surrounding predictions with a higher confidence level than itself while ignoring the predictions with a lower confidence level.

In the spatial adjacency relation M^N, each prediction only focuses on the prediction above the IOU threshold I_{thresh} with itself. Predictions too far away from each other generally do not correspond to the same GT, so setting the threshold I_{thresh} avoids inefficient computation. The right of Figure 4 shows how M^N and M^A obtain the final asymmetric relationship. For each query on the horizontal axis, the vertical axis is the visible relationship of other queries in the mask to that query.

Intuitively, in the asymmetric relationship, the predictions whose confidence score is between C_{low} and C_{high} only need to focus on the surrounding predictors with higher confidence scores than themselves and calculate the relative position relation. The field of view of each prediction in the asymmetric relationship is shown in Figure 5; in Figure 5a, the prediction box 2 can only see the information of the prediction box 1 with a higher confidence score than itself, and there are no repeat predictions in its field of view. In Figure 5b, the prediction box 3 can only see the information of the prediction boxes 1 and 2 and there are no repeat predictions in its field of view either. In Figure 5c, the prediction box 4 can see the information of the prediction boxes 1–3 with a higher confidence score than itself to judge itself as a duplicate prediction.



(a) **(b)** **(c)**

Figure 5. The field of view of each prediction in the asymmetric relationship. The dashed box represents the invisible box in the field of view; the brighter box color represents its higher confidence score; the confidence score ranking of the prediction boxes is 1 > 2 > 3 > 4.

3.2. Relation encoding

The asymmetric relation module $\mathcal{AR}$ accepts the set of attribute k consisting of N queries $P^k = \{k_0, k_1, \ldots, k_N\}$, and it computes the asymmetrical relation vector r^k for each query for the attribute k according to the asymmetric relation M determined by Eqs (3.3) and (3.4). We define those queries that the i^{th} query q_i can perceive in the asymmetric relation as a set $\mathcal{N}(q_i) = \left\{q_j | m_{ij} = 1\right\}$. We use two different relative relation encoding methods for the two different attributes k; Figure 6(a),(b) show the two different encoding network structures, while attribute k is the location of bounding boxes or of attention points sets, respectively.

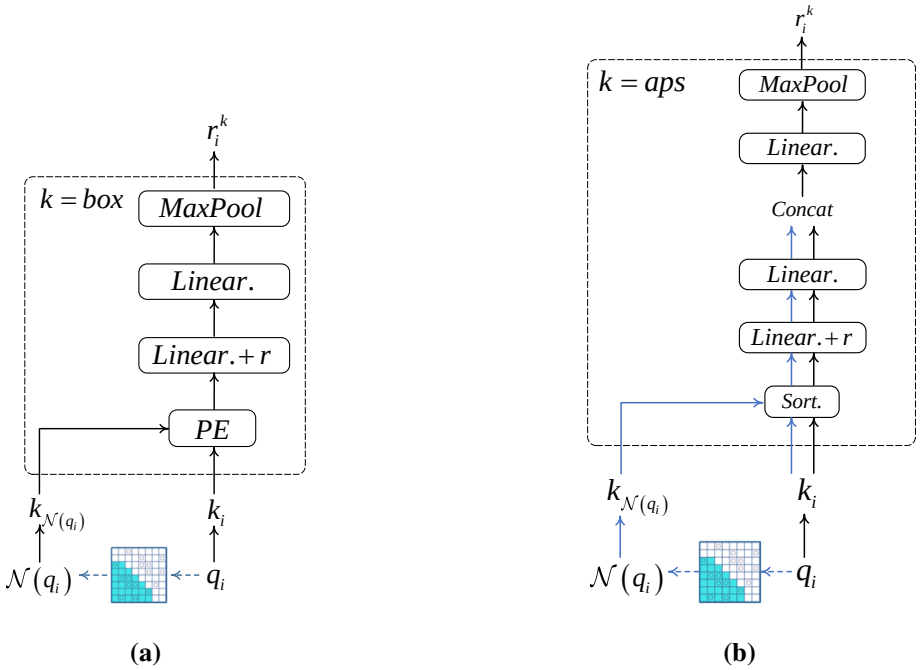


Figure 6. The figure shows the different encoding methods used in the ARN for the different relations k. Figure 6a is the encoding network when k is the relative position relation between the prediction boxes. Figure 6b is the encoding network when k is the relative relation between sets of attention points of the predictions. *PE* - sine and cosine spatial positional encoding function [24], r - RELU, *Linear* -fc layer, *Concat* - concatenate.

In Figure 6a, when attribute k is the location of bounding boxes, we reference the encoding method in E2EDet [23] to encode the relative relation of bounding boxes. $\mathcal{AR}^{box}$ can be formulated as in Eq (3.5).

$$\begin{aligned} r_{ij}^{box} &\leftarrow \mathcal{H}\left(PE(box_i - box_j, iou_{ij})\right) \\ r_i^{box} &\leftarrow MaxPool(\left\{r_{ij}^{box} | m_{ij} = 1\right\}) \end{aligned} \tag{3.5}$$

where $PE(\cdot)$ refers to the sine and cosine spatial positional encoding function [24], and $\mathcal{H}(\cdot)$ represents the function used to encode relative position relations, implemented by two fully-connected layers. r_{ij}^{box} denotes the relative relation encoding vector between the i^{th} query and the j^{th} query.

For the i^{th} query, $\mathcal{AR}^{box}$ first accepts the length, width and center-point coordinates and the *iou* value of the predicted boxes at the last stage. Then, we calculate the difference between the length,

width and center point coordinates of the query prediction box in the $\mathcal{N}(q_i)$ set, and the difference is cosine-encoded. Finally, the relation encoding vectors, which are visible in the asymmetric relation M, are maxpooled to get the final asymmetrical relation vector r_i^{box} for the i^{th} query.

In Figure 6b, the attribute k is the attention point sets. Because many attention points lead to enormous space complexity of the sine and cosine spatial positional encoding [24], we propose a new encoding method to encode the relative relation of attention point sets that considers both the topology of the point set itself and the position relationship between point sets. $\mathcal{AR}^{aps}$ can be formulated as in Eq (3.6).

$$\begin{aligned} O\left(G_i, G_j\right) &\leftarrow \mathcal{H}\left(sort(\{pt_n - pt_m | n < m\})\right), p_n \in G_i, p_m \in G_j \\ r_{ij}^{aps} &\leftarrow fcn\left(concat\left(O\left(G_i, G_j\right), O\left(G_i, G_i\right)\right)\right) \\ r_i^{aps} &\leftarrow MaxPool(\left\{r_{ij}^{aps} | m_{ij} = 1\right\}) \end{aligned} \tag{3.6}$$

where G_i denotes the attention point set $\{pt_n\}_{n=0}^{T}$, which contains T attention points of the i^{th} query, and $sort(\cdot)$ refers to the ordering of the relative distances between attention points from small to large. It adds an ordered prior to the distance vector to accelerate convergence. $\mathcal{H}(\cdot)$ is the function used to encode relative position relations, which is implemented by two fully-connected layers. $O\left(G_i, G_j\right)$ encodes relative position relations between two attention point sets G_i and G_j. However, since the point sets themselves have different topologies, we use the relative position relation $O(G_i, G_i)$ with itself for the normalized correction. $concat(\cdot)$ denotes vector concatenation, $fcn(\cdot)$ denotes a fully connected layer and r_{ij}^{aps} is relative relation encoding vector between two attention point sets. $MaxPool(\cdot)$ is applied to relative relation encoding vectors which are visible in the asymmetric relation M to get the final asymmetrical relation vector r_i^{aps} for the i^{th} query.

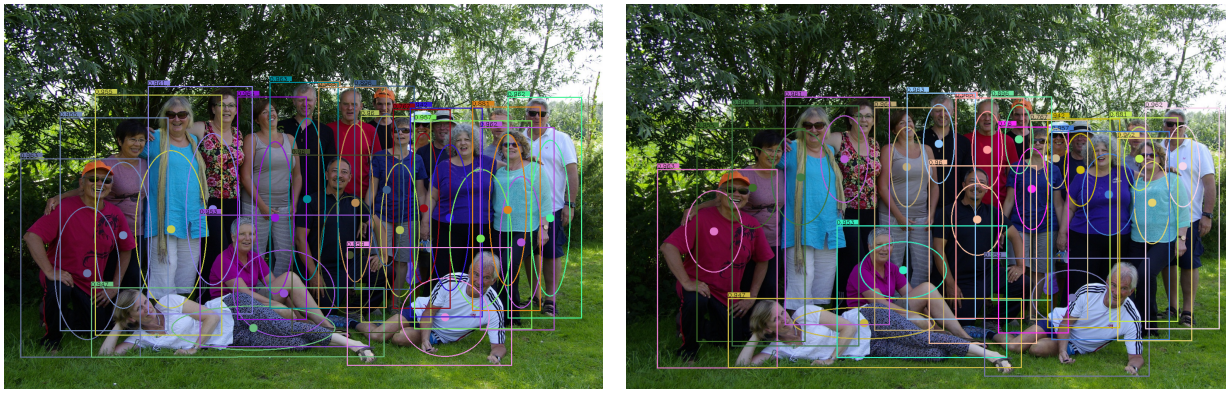
3.3. Decoupled cross attention head

We propose a decoupled cross-attention module, which can partially decouple the attention information required by the regression and category. We decoupled N_c out of eight attention heads, forcing the network only to select the attention points of N_c/N and the corresponding N_c/N attention heads to learn prediction confidence information. All original attention point information is still used for the position regression prediction of the prediction box. Our proposed DMCA(q, p_q, e_f) can be formulated as follows:

$$\begin{aligned} {q'}_t &= \sum_{n=1}^{N} W_n \left[\sum_{k=1}^{K} A_{nqk} \cdot {W_n}' e_f(p_q + \Delta p_{nqk})\right] \\ {q'}_t^c &= \sum_{n=N_c}^{N} W_n \left[\sum_{k=1}^{K} A_{nqk} \cdot {W_n}' e_f(p_q + \Delta p_{nqk})\right] \end{aligned} \tag{3.7}$$

where n indexes the attention head, k indexes the sampled keys and K is the total number of sampled keys. A_{nqk} and Δp_{nqk} denote the attention weight and the sampling offset of the k^{th} sampling point in the n^{th} attention head. p_q is the centroid coordinates of the prediction box corresponding to the query, which is used to get the attention point coordinates. $W_n \in C \times C_v (C_v = C/N)$ represents the learnable weights and A_{nqk} is the weight of attention points normalized by $\sum_{k=1}^{K} A_{nqk} = 1$. DMCA$(q, p_q, e_f)$

accepts the queries q, p_q from the previous stage and the keys and values from the encoder feature maps e_f, and it outputs the query $q_t'^c$ for confidence prediction and the query q'_t for location prediction, respectively.



(a) **(b)**

Figure 7. Attention point distribution. Figure 7a shows the attention point distribution from the original un-decoupled eight attention heads. Figure 7b shows the attention point distribution from the four attention heads decoupled out by *DMCA* to predict the class confidence. In Figure 7, each ellipse's long and short axes are the horizontal and vertical distributions' standard deviation of the attention points set, and each ellipse's center is the weighted average of the attention points.

Notably, the $q_t'^c$ and q'_t generated in Eq (3.7) use the same learnable parameters W_n without adding any additional parameter. $q_t'^c$ and q'_t are fed into the FFN with the same parameters to generate q_t^c and q_t. At the end of each decoder stage, the classification head C and the regression head $\mathcal{B}$ conduct a confidence prediction and a regression prediction using q_t^c and q_t for this stage, respectively.

Figure 7 visualizes the attention points distribution for the confidence prediction comparing the DMCA of $N_c = 4$ and $N_c/N = 0.5$ with the original muti-head cross-attention. Using DMCA, the network learns to predict confidence by selecting only a small number of attention points, so the network will be forced to focus more on regions with a higher confidence contribution, such as visible regions or human heads. The method can avoid the noisy information from surrounding GT-obscured regions to reduce false predictions. Figure 8 visualizes other examples of the confidence prediction attention point distributions decoupled by DMCA.

4. Experiment

In this section, we discuss the experiments on the CrowdHuman [17] dataset. We analyze the detection results for different crowded scenes and verify the effectiveness and robustness of our method. At the same time, we conducted ablation experiments on our proposed modules and their hyperparameters.

Figure 8. Distribution of attention points decoupled by our proposed *DMCA* to predict the class confidence. Each ellipse's long and short axes are the horizontal and vertical distributions' standard deviation of the attention point set, and each ellipse's center is the weighted average of the attention points.

4.1. Evaluation metric

We mainly take three criteria: average precision (AP), log-average miss rate MR^{-2} [25] and Jaccard index (JI) [26] as evaluation metrics. Generally, good detection results correspond to a larger AP, a larger JI and a smaller MR^{-2}.

- The AP is represented by the area surrounded by the precision-recall curve and coordinates. AP is commonly used in object detection to reflect both precision and recall, and a larger AP indicates better performance.
- MR^{-2} [25] computes the average miss rate on a log scale of false positives per image. This metric is commonly used in pedestrian detection as it reflects false and missing detection, and a smaller MR^{-2} indicates better performance.
- The JI [26] mainly evaluates how much the prediction set overlaps the GTs. It reflects the total distribution's similarity between the prediction box set and the real GTs. A larger JI indicates better performance.

4.2. Detailed settings

We used a standard ResNet-50 [27] pre-trained on ImageNet [28] as the backbone of deformable DETR, and each training ran for 55 epochs. We trained our model with the AdamW optimizer; the momentum was 0.9, and the weight decay was 0.0001. The model's learning rate was 0.0002, and the backbone's learning rate was 0.00002. The batch size was 8, and the task was split into four GPUs. If there are no special instructions, the query number was set to 1000, the attention head's number was 8

Table 1. Comparative experimental results on CrowdHuman validation set.

Method	#Queries	AP ↑	MR^{-2} ↓	JI ↑
Box-based				
RetinaNet [29]	-	85.3	55.1	73.7
ATSS [30]	-	87.0	55.1	75.9
ATSS [30] + MIP [4]	-	88.7	51.6	77.0
Faster-RCNN [16]	-	85.0	50.4	-
FPN [5]+Adaptive-NMS [6]	-	84.7	47.7	-
FPN [5]+Soft-NMS [7]	-	88.2	42.9	79.8
FPN [5]+MIP [4]	-	90.7	41.4	82.3
PBM [19]	-	89.3	43.3	-
Query-based				
DETR [11]	100	75.9	73.2	74.4
PED [22]	1000	91.6	43.7	83.3
Sparse-RCNN [12]	500	90.7	44.7	81.4
Sparse-RCNN [12]	750	91.3	44.8	81.3
D-DETR [9]	1000	91.3	43.8	83.3
Iter-E2EDet [23]	1000	92.1	41.5	84.0
D-DETR+ours (1 lay)	1000	**92.6**	40.0	**84.4**
D-DETR+ours (2 lay)	1000	92.5	**39.7**	84.3

*Note: where N lay denotes the number of asymmetrical layers.

in the deformable DETR and the block number for both the encoder and decoder was set to 6. For a fair comparison, we also used six decoders divided into five decoupled decoders and one asymmetrical relation decoder; the other settings of the six decoders were the same for the deformable DETR. We selected the best performance in the last five epochs for recording in each training set.

4.3. Comparative experiment on CrowdHuman

The CrowdHuman [17] dataset contains 15,000, 4,370 and 5,000 images for training, validation and testing, respectively. For a fair comparison, we conducted all experiments on the validation set using full-body annotations in the same environment. We conducted comparative experiments with mainstream detectors, including box-based detectors [4–7, 16, 19, 29] and query-based detectors [9, 11, 12, 22, 23]. Table 1 shows the experimental results.

We find that the end-to-end query-based methods generally perform better than the box-based methods based on greedy algorithm post-processing in crowded scenes. Surprisingly, PEDR [22], designed for crowded scenes and pedestrian detection based on deformable DETR [9], had very limited performance improvement on the CrowdHuman dataset. When using a sparse number of queries, the end-to-end CNN-based approach sparse-RCNN [12] was highly competitive and still performed well with only 500 queries. Using our proposed deformable DETR, the model achieved a 92.6% AP, 40.0% MR^{-2} and 84.4% JI with 1000 queries. Compared with the baseline deformable DETR [9], we improved by 1.3% for the AP, 3.8% for the MR^{-2} and 1.1% for the JI. Compared with the past best results of Iter-E2EDet [23], we improved by 0.5% for the AP, 1.5% for the MR^{-2} and

0.4% for the JI. Our improved model demonstrated significant improvement in all three metrics.

4.4. Ablation study of different module

To verify the effectiveness of our proposed asymmetrical relation module *AR* and *DMCA* in the previous Section 3, we performed ablation experiments based on the deformable DETR with a backbone of R50 and 1000 queries. Table 2 shows the ablation experiments of different modules. Comparing with the baseline model [9], our asymmetric relation module *AR* had the following improvements: 1.2% AP, 2.4% MR^{-2} and 0.9% JI, and our decoupled attention mechanism *DMCA* can mainly improve by another 0.4% MR^{-2}. Figure 9 shows our method's intermediate results from different decoder stages and the original deformable DETR. In our method, the last asymmetrical relation decoder with the *AR* module can significantly increase AP and decrease MR^{-2}.

Table 2. Ablation experiments on *CrowdHuman* validation set.

AR	DMCA	AP ↑	MR^{-2}↓	JI ↑
		91.3	43.8	83.3
	✓	91.7	42.9	83.4
✓		92.5	40.4	84.2
✓	✓	**92.6**	**40.0**	**84.4**

*Note: The baseline model (the first line) is deformable DETR [9] with ResNet-50 [27]. *AR*—asymmetrical relation module. *DMCA*—decoupled mult-dead cross-attention.

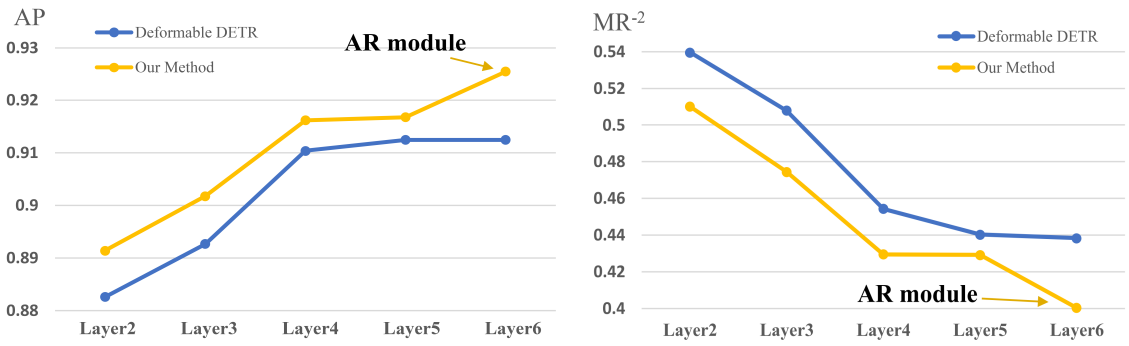


Figure 9. Detection result comparison for different stage decoders between our method and the original deformable DETR.

4.5. Analysis of asymmetric relation module

To analyze the effect of the asymmetric relation mask determined by Eqs (3.3) and (3.4) in the asymmetric relation module, and also to verify the effectiveness of the asymmetric relation module, we conducted comparative experiments. When the decoupled attention N_c was fixed at 4; we used a relationship module in the last layer to change the values of C_{high} and C_{low} under the same conditions for comparative experiments; the results are presented in Table 3. When C_{high} is fixed at 0.7, C_{low} varies from 0.05 to 0.4. When C_{low} is 0.1, our method has the best overall performance. When C_{low} is fixed at 0.1, C_{high} varies from 0.7 to 1. When C_{high} is 1.0, our method has the best overall performance.

The experimental results show that a larger range of asymmetric relation can improve the detector performance.

Table 3. Experimets with different values of C_{low} and C_{high} in the asymmetric relation module.

C_{low}	C_{high}	AP ↑	MR^{-2}↓	JI ↑
0.05	0.70	92.5	41.6	84.2
0.10	0.70	92.4	40.9	**84.7**
0.20	0.70	92.3	41.4	84.4
0.30	0.70	92.3	41.2	84.1
0.40	0.70	92.4	41.6	84.4
0.10	0.80	92.5	41.1	84.2
0.10	0.90	92.5	**39.9**	84.1
0.10	1.00	**92.6**	40.0	84.4

4.6. Analysis of decoupled cross attention head

To analyze the impact of different proportions of points decoupled by DMCA, we conducted experiments by varying different values of N_c when C_{low} is set to 0.1 and C_{high} is set to 1.0. Our model can decouple different proportions N_c/N of attention points for confidence prediction. The experimental results are shown in Table 4. Our model works best when N_c/N is 0.5, meaning that half of the attention points are taken to predict confidence.

Table 4. Performance of the DMCA with different hyperparameters N_c.

N_c	N_c/N	AP ↑	MR^{-2}↓	JI ↑
2	0.25	**92.6**	40.4	84.4
3	0.375	92.4	40.8	84.3
4	0.5	**92.6**	**40.0**	**84.4**
5	0.625	92.5	40.1	84.3
6	0.75	92.5	40.7	84.2

4.7. Ablation study of different attributes k in the asymmetric relation module

To verify the validity of different encoding methods corresponding to different attributes k in Figure 6, we experimented with two different encodings of relations k in Iter-E2EDet [23] and our method, respectively. In Table 5, *Box* is the relative relation of the predicted bounding boxes, and Aps is the relative relation between attention point sets. The asymmetric networks of both relations have a significant improvement over the baseline, proving the effective competitiveness of the coding method for the relative relation of attention point sets.

4.8. Analysis of different number of queries and AR layers

In this section, we analyze the effects of different numbers of asymmetric blocks with asymmetric relations and different numbers of queries. Table 6 shows the experimental results. We can find that

Table 5. Ablation experiments of different coding relations k.

Method	*Box*	*Aps*	AP ↑	MR^{-2}↓	JI ↑
deformable DETR [9]			91.3	43.8	83.3
Iter-E2EDet [23]	✓		92.1	**41.5**	84.0
		✓	**92.1**	41.6	**84.3**
Our Method	✓		**92.6**	**40.0**	**84.4**
		✓	92.5	40.2	84.2

model performance does not improve significantly as the number of asymmetrical decoders with asymmetric relation increases. However, while we use the asymmetric relation module only in the last layer, it is the best configuration. Also, we can find a significant improvement in AP after increasing the number of queries, but with a reduction in MR^{-2}, probably because more queries are needed to match the GTs in crowded scenes.

Table 6. Performance comparison for different numbers of queries and asymmetrical decoders.

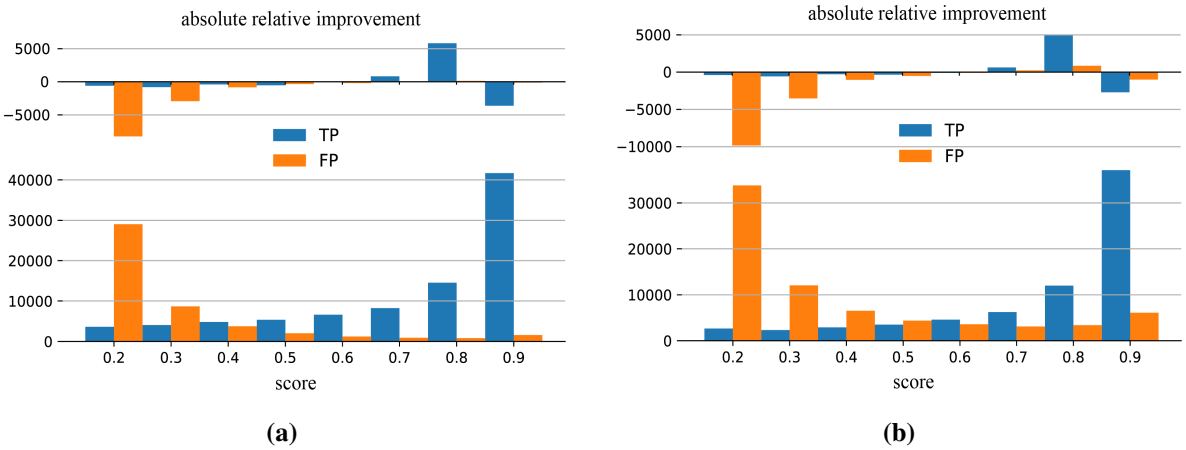
Number of *AR* layers	#Queries	AP ↑	MR^{-2}↓	JI ↑
1	1000	92.6	40.0	84.4
2	1000	92.5	**39.7**	84.3
3	1000	92.3	40.5	84.1
1	2000	**92.9**	40.7	84.3
2	2000	**92.9**	40.8	**84.5**

4.9. Analysis of false positves in different scenes

To analyze the enhancement of our method in detail, we counted the false positives (FPs) and true positives (TPs) of our method and the baseline deformable DETR at different confidence scores. In Figure 10, we show the statistical results of FPs and TPs of the deformable DETR at different confidence scores, and the relative improvement of our method when the matching rule is at the IOU threshold 0.5 and IOU threshold 0.7, respectively. Our method can reduce many FPs at a low confidence range and increase more TPs at a high confidence range.

To verify the effectiveness and robustness of our method for different crowded scenes, we counted the FPs and TPs under different crowded scenes and their relative variations compared to the baseline. We define the crowdedness of each GT as the maximum value of the the IOU between that GT and its surrounding GTs. We divide the GTs by different crowdedness and count their FP and TP separately. Figure 11 shows the statistics of TPs and FPs for different crowded scenes of the baseline deformable DETR and the relative improvement of our method when the matching rules of FP and TP are IOU > 0.5 and confidence score > 0.7. The results show that our method can significantly reduce FPs and increase TPs for different crowded scenes, verifying our proposed method's robustness against the crowdedness.

Figure 12 compares the detection results between our method and the baseline deformable DETR. Our method still has high recall and significantly fewer duplications and missed predictions when GTs



(a) **(b)**

Figure 10. Comparison of our method and deformable DETR for FP and TP statistical results at different confidence scores when the IOU matching threshold is set to 0.5 in Figure 10a and 0.7 in Figure 10b. The bottom histogram for each figure describes the prediction distribution of the baseline deformable DETR [9] for different confidence scores, while the top one reflects the relative improvements achieved by our approach as compared with the baseline.

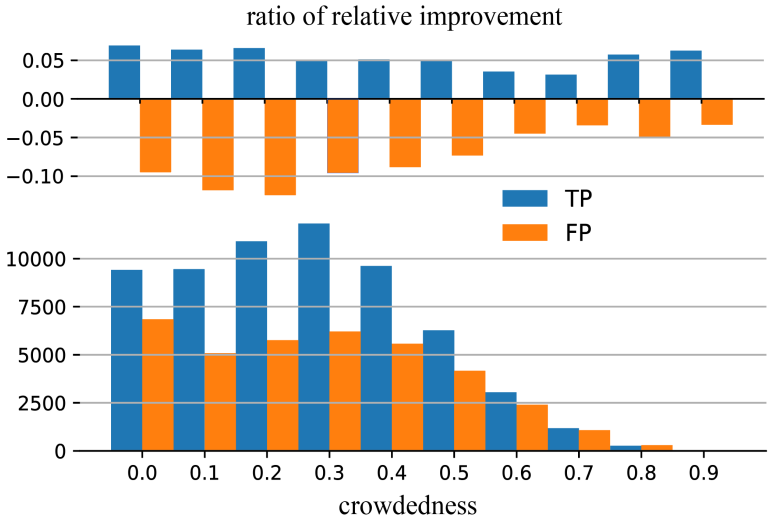


Figure 11. The bottom histogram describes the prediction distribution of the deformable DETR [9] for different crowded scenes, while the top one reflects the relative improvements achieved by our method as compared with the deformable DETR.

Figure 12. Result visualization for the baseline deformable DETR (above) and our method (below). Only predictions with a confidence score higher than 0.3 are plotted in the figure.

are heavily occluded or are just small visible areas in crowded scenes.

4.10. Performance of our method with large model in crowded scenes

To explore the detection upper bound of our method under the condition of crowded scenes, we replaced the ResNet50 with a large backbone, i.e., Swin-Large [31]. Experiments were conducted with the same training strategy as described in Section 4.2; our method obtained **94.3%** AP, **36.3%** MR^{-2} and **87.4%** JI with 1000 queries, which constitutes a state-of-the-art result on CrowdHuman validating datasets.

5. Conclusions

In this paper, we propose asymmetric relational network modules and decoupled cross-attention heads to improve the performance of query-based models on crowded scene data. Using our approach, the deformable DETR improves the de-duplication ability and reduces the miss rate of prediction, with stable performance gains for different crowded scenes.

Since our approach has been implemented on a DETR-like model, which requires a large amount of computing resources, our next step is to study how to reduce the model parameters. Moreover, our asymmetric relation module is only implemented in the last several stages of the fine-tuning process; how to implement asymmetric relations in the whole fine-tuning stage still needs to be explored.

Use of AI tools declaration

The authors declare that they have not used artificial intelligence tools in the creation of this article.

Acknowledgments

This research was funded by the Natural Science Foundation of China (grant no. 62162065, 62061049, 12263008), the Application and Foundation Project of Yunnan Province (grant no. 202001BB050032), the Department of Science and Technology of Yunnan Province–Yunnan University Joint Special Project for Double-Class Construction (grant no. 202201BF070001-005), the Expert Workstation of Yunnan Province (202105AF150011) and the Postgraduate Practice and Innovation Project of Yunnan University (grant no. 22221264).

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016), 779–788. https://doi.org/10.1109/CVPR.2016.91
2. C. Fu, W. Liu, A. Ranga, A. Tyagi, A. C. Berg, DSSD: Deconvolutional single shot detector, preprint, arXiv: 1701.06659.
3. J. Redmon, A. Farhadi, Yolov3: An incremental improvement, preprint, arXiv: 1804.02767.
4. X. Chu, A. Zheng, X. Zhang, J. Sun, Detection in crowded scenes: One proposal, multiple predictions, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 12211–12220. https://doi.org/10.1109/CVPR42600.2020.01223
5. T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, S. J. Belongie, Feature pyramid networks for object detection, in *2017 IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 936–944, https://doi.org/10.1109/CVPR.2017.106
6. S. Liu, D. Huang, Y. Wang, Adaptive NMS: refining pedestrian detection in a crowd, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2019), 6459–6468, https://doi.org/10.1109/CVPR.2019.00662
7. N. Bodla, B. Singh, R. Chellappa, L. S. Davis, Soft-nms - improving object detection with one line of code, in *IEEE International Conference on Computer Vision*, (2017), 5562–5570. https://doi.org/10.1109/ICCV.2017.593
8. S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, et al., DAB-DETR: dynamic anchor boxes are better queries for DETR, in *The Tenth International Conference on Learning Representations*, 2022.
9. X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable DETR: deformable transformers for end-to-end object detection, preprint, arXiv: 2010.04159.
10. F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, L. Zhang, DN-DETR: accelerate DETR training by introducing query denoising, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 13609–13617. https://doi.org/10.1109/CVPR52688.2022.01325

11. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in *Computer Vision – ECCV 2020: 16th European Conference*, (2020), 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

12. P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, et al., Sparse R-CNN: end-to-end object detection with learnable proposals, in *IEEE Conference on Computer Vision and Pattern Recognition*, (2021), 14454–14463. https://doi.org/10.1109/CVPR46437.2021.01422

13. H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, et al., DINO: DETR with improved denoising anchor boxes for end-to-end object detection, preprint, arXiv: 2203.03605.

14. T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, et al., Microsoft COCO: common objects in context, in *Computer Vision - ECCV 2014 - 13th European Conference*, (2014), 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

15. M. Everingham, S. M. Eslami, L. V. Gool, C. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: A retrospective, *Int. J. Comput. Vision*, **111** (2015), 98–136. https://doi.org/10.1007/s11263-014-0733-5 .

16. S. Ren, K. He, R. B. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, *IEEE Trans. Pattern Anal. Mach. Intell.*, **39** (2017), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

17. S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, et al., Crowdhuman: A benchmark for detecting human in a crowd, preprint, arXiv:1805.00123.

18. C. Zhou, J. Yuan, Bi-box regression for pedestrian detection and occlusion estimation, in *Computer Vision - ECCV 2018 - 15th European Conference*, (2018), 138–154. https://doi.org/10.1007/978-3-030-01246-5_9

19. X. Huang, Z. Ge, Z. Jie, O. Yoshie, NMS by representative region: Towards crowded pedestrian detection by proposal pairing, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 10747–10756. https://doi.org/10.1109/CVPR42600.2020.01076

20. S. Zhang, J. Yang, B. Schiele, Occluded pedestrian detection through guided attention in cnns, in *2018 IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 6995–7003. https://doi.org/10.1109/CVPR.2018.00731

21. Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, L. Shao, Mask-guided attention network for occluded pedestrian detection, in *2019 IEEE/CVF International Conference on Computer Vision*, (2019), 4966–4974. https://doi.org/10.1109/ICCV.2019.00507

22. M. Lin, C. Li, X. Bu, M. Sun, C. Lin, J. Yan, et al., Detr for crowd pedestrian detection, preprint, arXiv: 2012.06785.

23. A. Zheng, Y. Zhang, X. Zhang, X. Qi, J. Sun, Progressive end-to-end object detection in crowded scenes, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2022), 847–856. https://doi.org/10.1109/CVPR52688.2022.00093

24. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al., Attention is all you need, in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, (2017), 5998–6008.

25. P. Dollár, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: An evaluation of the state of the art, *IEEE Trans. Pattern Anal. Mach. Intell.*, **34** (2012), 743–761. https://doi.org/10.1109/TPAMI.2011.155

26. W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, et al., SSD: single shot multibox detector, in *Computer Vision - ECCV 2016 - 14th European Conference*, (2016), 21–37. https://doi.org/10.1007/978-3-319-46448-0_2

27. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *2016 IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778. https://doi.org/10.1109/CVPR.2016.90

28. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, et al., Imagenet large scale visual recognition challenge, *Int. J. Comput. Vis.*, **115** (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

29. T. Lin, P. Goyal, R. B. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in *IEEE International Conference on Computer Vision*, (2017), 2999–3007. https://doi.org/10.1109/ICCV.2017.324

30. S. Zhang, C. Chi, Y. Yao, Z. Lei, S. Z. Li, Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2020), 9756–9765. https://doi.org/10.1109/CVPR42600.2020.00978

31. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision*, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

AIMS Press