

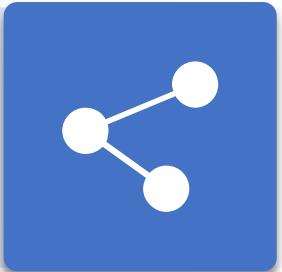


# Progressive End-to-End Object Detection in Crowded Scenes



Anlin Zheng, Yuang Zhang, Xiangyu Zhang, Xiaojuan Qi, Jian Sun  
Accepted by CVPR 2022

智能网络与优化实验室





1

# Introduction



## 介绍

拥挤对象检测 (Crowded Object Detection) 是计算机视觉中一个实用且具有挑战性的研究领域。对于端到端的目标检测框架 DETR，引入了可学习的 query 来表示对象，并在没有任何后处理 (post-processing) 的情况下取得了较好的性能，这可以归类为 query-based 的方法，区别于 box-based 和 point-based 的方法。而 Deformable DETR 提出了关注一部分特征点的注意模块，提高了检测精度，但也带来了收敛速度慢、计算开销大的问题。

这些 query-based 的方法在 COCO 等稍微拥挤的场景中获得了显著的结果，但作者研究发现，这些方法在拥挤场景中依然存在为解决的挑战：

- query-based 的方法倾向于预测单个目标的多个预测，  
并引入 false-positive 机制。
- 随着解码阶段深度的增加，基于查询的检测器的性能变得饱和甚至更差。

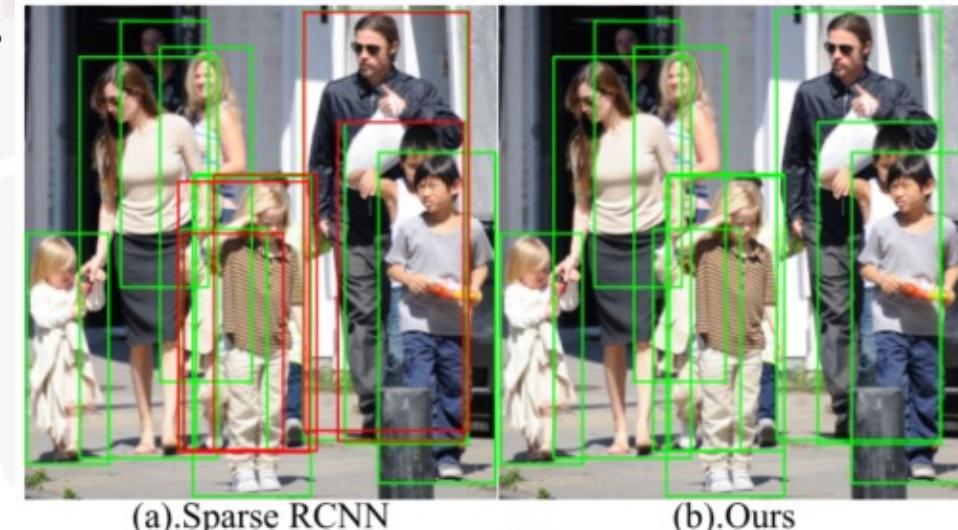


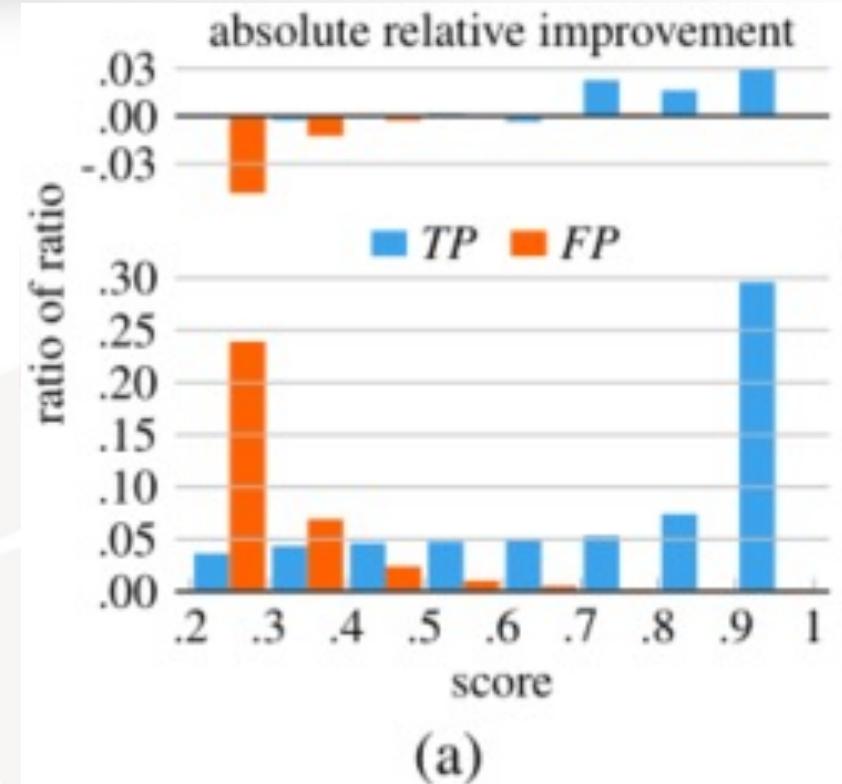
Figure 1: 1a. Sparse RCNN [39] introduces *false positives* in crowded scenes. 1b. Our approach can remove those *false positives* and ensure each object can be detected only once. Green boxes indicate *true positives* while red ones represent *false positives*.



## 动机

进一步研究 query-based 的方法，如 Sparse R-CNN，如图，那些具有高置信度分数（如高于 0.7）的预测可以准确预测出很多的目标对象，同时包含很少的误报。这些，可以被认为是可接受的预测。

而其余的，存在相当数量的 true positive 和 false positive，可以被视为 noisy predictions。当然，如果目标被一个 accepted prediction 检测到，就不需要再用噪声预测来检测它。因此，在被接受的预测的情况下，可以加强对那些嘈杂预测的区分。为此，noisy queries 要可以“感知”它们的目标是否被检测到。



(a)

Figure 2: 2a. The bottom histogram describes the prediction distribution of Sparse RCNN [39] under different confidence scores, while the top one reflects the absolute improvements achieved by our approach compared with Sparse RCNN [39]. 2b. The FP-TP curve when computing



## 贡献

提出了一种配备预测选择器、关系信息提取器、查询更新器和标签分配的渐进式预测方法，以提高基于查询的对象检测器在处理拥挤场景时的性能。

- 开发了一个预测选择器来选择有高置信度分数的 query 作为 accepted queries，其他作为 noisy queries。
- 为了让 noisy queries “感知” 它们的目标是否已被检测到，设计了一个关系提取器，用于 noisy queries 与其可接受的邻居之间的关系建模。
- 查询更新器通过执行一种新的局部自注意力，只关注与空间相关的邻居来进行开发。
- 引入了新的一对一标签分配规则，逐步在接受（accepted）和细化（refined）的 noisy queries 中分配样本。



2

# Methodology



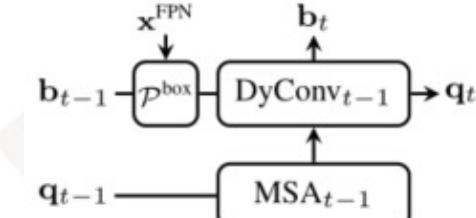
## Query Based 的目标检测

本文的方法可以用于大多数 query-based 的目标检测器，此处选择了 Sparse R-CNN 作为默认的实例。Sparse R-CNN 的 pipeline 如图，可以用式子描述为：

$$\begin{aligned} x_{t-1} &\leftarrow \mathcal{P}^{box}(x^{FPN}, b_{t-1}), & q_{t-1}^* &\leftarrow \text{MSA}_{t-1}(q_{t-1}) \\ q_t &\leftarrow \text{DynConv}_{t-1}(q_{t-1}^*, x_{t-1}), & b_t &\leftarrow \mathcal{B}_{t-1}(q_t) \end{aligned}$$

-  $q \in R^{N \times d}$  表示科学其的目标查询， $N$  和  $d$  表示查询  $q$  的数量和维度。

- 在阶段  $t$ ，在前一阶段预测的边界框  $b_{t-1}$  的引导下，RoI Align 过程  $\mathcal{P}^{box}$  从 FPN 特征  $x^{FPN}$  中提取 RoI 特征。
- 多头注意力机制  $\text{MSA}_{t-1}$  作用于输出查询  $q_{t-1}$  产生变换后的查询  $q_{t-1}^*$ 。
- 动态卷积模块  $\text{DynConv}_{t-1}$  对  $q_{t-1}^*$  和  $x_{t-1}$  进行动态卷积来为下一阶段产生查询  $q_t$ 。同时， $q_t$  也被放入 box 预测分支  $\mathcal{B}_{t-1}$  用于产生当先的边界框预测  $b_t$  并作为下一阶段的输入。



(a) Sparse R-CNN [39].



## 本文的方法

如图，本文提出的方法由预测选择器 ( prediction selector )、关系信息提取器 ( relation information extractor )、查询更新器 ( query updater ) 和标签分配 ( label assignment ) 几个部分组成。

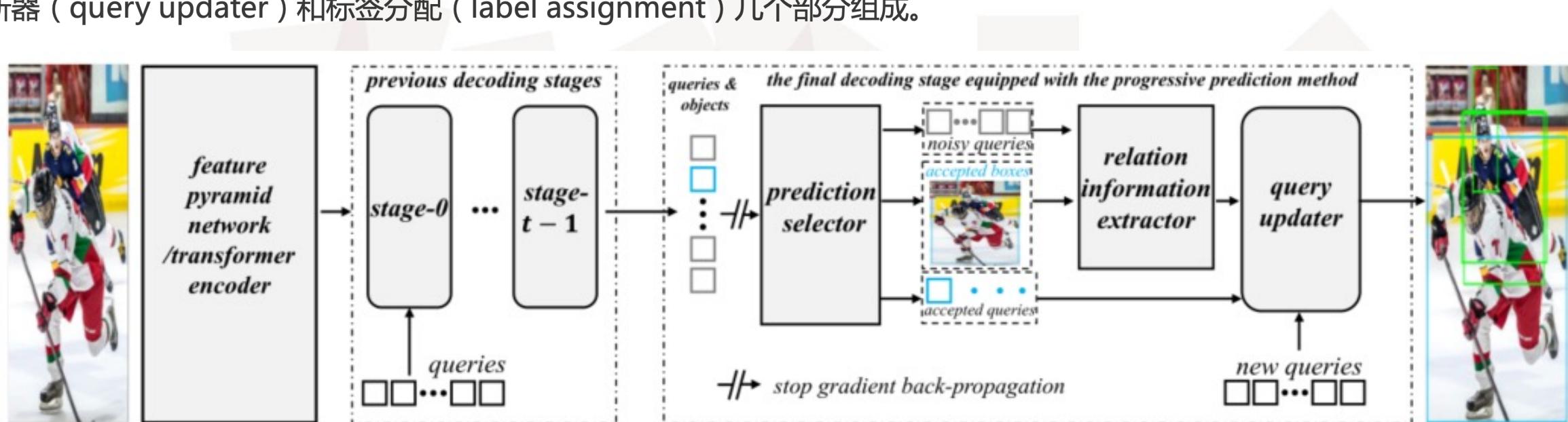


Figure 3: The diagram of the proposed progressive end-to-end object detection framework. First, the *prediction selector* select queries associated with high confidence scores as *accepted queries*, leaving the rest as *noisy queries*. Then *Relation information extractor* models the relations between *noisy queries* and their neighbors from accepted predictions. Next, the queries are fed into the *queries updater* to be further refined by performing a new local self-attention.



## 预测选择器 ( Prediction Selector )

预测选择器用于将那些容易生成具有高置信度预测的 query 作为 accepted queries , 而将其余的 query 保留位需要进一步细化分析的 noisy queries。成过程表达为下式 :

$$\mathcal{D}_{t-1}^h \leftarrow \{b_i | s_i \geq s \wedge b_i \in \mathcal{D}_{t-1}\}, \quad \mathcal{D}_{t-1}^l \leftarrow \mathcal{D}_{t-1} - \mathcal{D}_{t-1}^h$$

其中  $t$  为阶段数 ,  $\mathcal{D}_{t-1}$  为在  $t-1$  阶段由所有的 query 产生的所有预测 ,  $\mathcal{D}_{t-1}^h$  和  $\mathcal{D}_{t-1}^l$  表示由 accepted queries 和 noisy queries 产生的可接受的预测和噪声预测 ,  $b_i$  和  $s_i$  表示预测框及其置信度 ,  $s$  为置信度阈值。



## 关系信息提取器 ( Relation Information Extractor )

前面提到，大部分的 target 目标可以被从 accepted queries 中精确地预测出来。因此，如果一个目标可以被一个可接受的预测所检测出来，那么就没必要用噪声预测去再检测一次。关系信息提取器则用于使 noisy queries 具有感知目标是否被检测到的能力。关系信息提取器的设计如图，可以由下面的关系式所表示。

对于每一个噪声预测  $b_i$ ，首先在  $\mathcal{D}_{t-1}^h$  找到它们的可接受邻居  $\mathcal{N}(b_i)$ ，构造关系对  $(b_i, \mathcal{N}(b_i))$ 。然后被编码的关系对与关系对产生的 IoU 被送入网络中产生几何关系特征  $\mathcal{H}(b_i)$ 。由于与每一个噪声预测相关联的可接受邻居数目是不确定的，因此使用一个聚合函数来作用于  $\mathcal{H}(b_i)$  以产生相同的特征维度，同时维护可接受邻居的排列不变性。几何特征与转换后的 query 特征融合在一起，通过非线性函数进一步激活。

$$\mathcal{N}(b_i) \leftarrow \{b_j | \mathcal{O}(b_i, b_j) \geq \theta\}, b_i \in \mathcal{D}_{t-1}^l, b_j \in \mathcal{D}_{t-1}^h$$

$$\mathcal{H}(b_i) \leftarrow \mathcal{U}(\mathcal{E}(b_i, \mathcal{N}(b_i))), b_i \in \mathcal{D}_{t-1}^l$$

$$\mathcal{R}(b_i) \leftarrow \mathcal{T}(\text{MaxPool}(\mathcal{H}(b_i)) + \mathcal{F}(q_i))$$

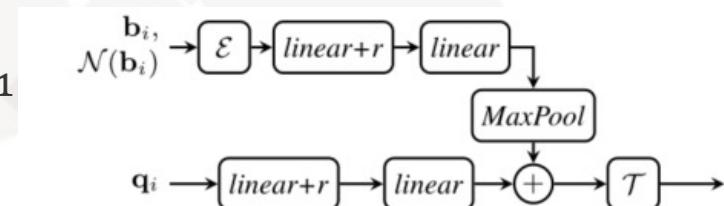


Figure 5: Relation information extractor  $\mathcal{R}$ .  $\mathcal{E}$  – sine and cosine spatial positional encoding function [20, 45], linear – fc layer,  $r$  – ReLU,  $\mathcal{T}$  – fc layer.



## 关系信息提取器 ( Relation Information Extractor )

$$\mathcal{N}(b_i) \leftarrow \{b_j | \mathcal{O}(b_i, b_j) \geq \theta\}, b_i \in \mathcal{D}_{t-1}^l, b_j \in \mathcal{D}_{t-1}^h$$

$$\mathcal{H}(b_i) \leftarrow \mathcal{U}\left(\mathcal{E}(b_i, \mathcal{N}(b_i))\right), b_i \in \mathcal{D}_{t-1}^l$$

$$\mathcal{R}(b_i) \leftarrow \mathcal{T}(\text{MaxPool}(\mathcal{H}(b_i)) + \mathcal{F}(q_i))$$

其中， $\mathcal{N}(\cdot)$  为基于 IoU  $\mathcal{O}(\cdot, \cdot)$  和阈值  $\theta$ ，在  $\mathcal{D}_{t-1}^h$  中找到  $\mathcal{D}_{t-1}^l$  中噪声预测  $b_i$  的可接受邻居的函数。 $\mathcal{E}(\cdot, \cdot)$  为 sine 和 cosine 空间位置编码函数。 $\mathcal{U}(\cdot, \cdot)$  表示产生几何特征  $\mathcal{H}(b_i)$  的函数。于  $b_i$  相关联的 noisy query  $q_i$  通过函数  $\mathcal{F}(\cdot)$  进行转换，然后与  $\mathcal{H}(b_i)$  通过 element-wise 求和后传入  $\mathcal{T}$  以产生期望的关系特征  $\mathcal{R}(b_i)$ 。

如上图， $\mathcal{U}(\cdot, \cdot)$  由两个全连接层和 ReLU 构成， $\mathcal{F}(\cdot)$  和  $\mathcal{U}(\cdot, \cdot)$  有相同的结构，但是权重不共享。

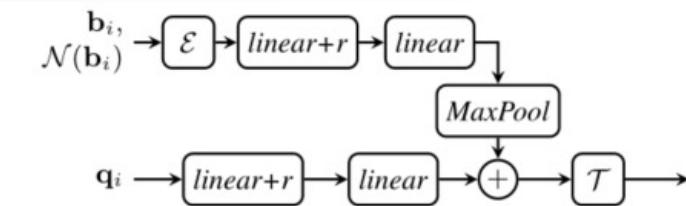


Figure 5: Relation information extractor  $\mathcal{R}$ .  $\mathcal{E}$  – sine and cosine spatial positional encoding function [20, 45], linear – fc layer,  $r$  – ReLU,  $\mathcal{T}$  – fc layer.



## 查询更新器 (Query Updater)

查询更新器用于进一步细化 noisy queries 的特征，可以表示为下式。由于  $\mathcal{D}_{t-1}^h$  和  $\mathcal{D}_{t-1}^l$  的分布不同于  $\mathcal{D}_{t-1}$ ，因此先用一组新的可学习 queries 通过元素求和的方式来补充关系特征。然后，被补充的 noisy queries 集合作为输入查询  $q_{t-1}$  执行局部自注意力 LMSA<sub>t-1</sub> 以及动态卷积。

$$q_{t-1} \leftarrow \{\hat{q}_i | \hat{q}_i = \mathcal{R}(b_i) + e_i\}, b_i \in \mathcal{D}_{t-1}^l, e_i \in E$$

$$q_{t-1}^* \leftarrow \text{LMSA}_{t-1}(q_{t-1})$$

由于目标检测主要聚焦于图片的局部区域，因此设计 LMSA<sub>t-1</sub> 确保每一个 query 只与局部邻居交互。LMSA<sub>t-1</sub> 首先基于框的 IoU (值大于 0) 找到每个 query 的邻居，然后执行注意力机制中的 “qkv” 操作。至此，注意力机制仅在局部进行。



## 标签分配 (Label Assignment )

首先，将可接受预测  $\mathcal{D}_{t-1}^h$  与目标的 GT 集进行匹配。然后，移除已经被匹配的目标，考虑噪声预测  $\mathcal{D}_t^l$  与余下的 GT 的二分匹配，匹配过程描述为如下算法：

---

### Algorithm 1 Label Assignment for $\mathcal{D}_t^l$ .

---

**Input:**  $\mathcal{D}_t^l, \mathcal{D}_t^h, \mathcal{G}$ ;

- 1:  $\mathcal{D}_t^l$ : results of  $\mathcal{D}_{t-1}^l$  in Equ.(2) from stage  $t$ ;
- 2:  $\mathcal{D}_t^h$ : results of  $\mathcal{D}_{t-1}^h$  in Equ.(2) from stage  $t$ ;
- 3:  $\mathcal{G}$ : target boxes.

**Output:** The matched predictions  $\mathcal{M}_D^l$  and corresponding targets  $\mathcal{M}_G^l$  after assignment.

- 4: Compute matching costs  $\mathcal{C}_t^h$  between  $\mathcal{D}_t^h$  and  $\mathcal{G}$ ;
- 5:  $\mathcal{M}_G^h, \mathcal{M}_D^h = \text{HungarianMatch}(\mathcal{D}_t^h, \mathcal{G}, \mathcal{C}_t^h)$ ;
- 6:  $\mathcal{G}_t^l = \mathcal{G} - \mathcal{M}_G^h$ ;
- 7: Compute matching costs  $\mathcal{C}_t^l$  between  $\mathcal{D}_t^l$  and  $\mathcal{G}_t^l$ ;
- 8:  $\mathcal{M}_G^l, \mathcal{M}_D^l = \text{HungarianMatch}(\mathcal{D}_t^l, \mathcal{G}_t^l, \mathcal{C}_t^l)$ ;
- 9: **return**  $\mathcal{M}_G^l, \mathcal{M}_D^l$ ;

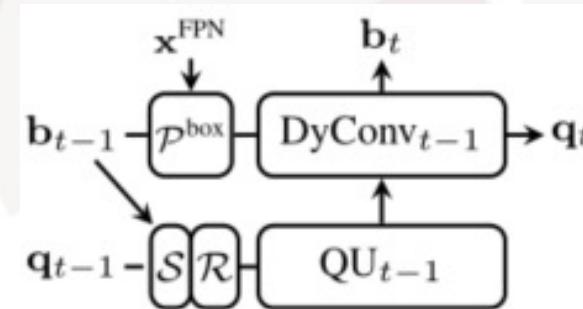


## 检测器设计上的特点

选用 Sparse R-CNN 作为实例，包含  $t$  ( $t = 6$ ) 个解码阶段组成，保持前  $t - 1$  个阶段不变，仅在最后阶段使用所提出的方法。

与 Sparse R-CNN 的主要不同点如下：

- 最后一层结构的变动，其结构变为如图。
- 前  $t - 1$  阶段使用 box 回归分支用于 box 预测，而最后一阶段直接使用前面  $t - 1$  阶段的结果用于训练和测试，而保留识别分支。
- 使用 Sparse R-CNN 和 Deformable DETR 中的集合预测损失。对于阶段  $t$ ，移除可接受预测  $\mathcal{D}_{t-1}^h$  中的样本。并为了缓解不平衡问题，尽早地拒绝置信度低于 0.05 的较明显的负样本。



(b) *Our approach*

MSA – Multi-head Self-Attention,  $S$  – Prediction Selector,  
 $\mathcal{R}$  – Relation Information Extractor, QU – Query Updater.



3

# Experiment



## 实验设置

- 数据集 : COCO , CrowdHuman , CityPersons
- 使用 momentum 为 0.9 , 权重衰减为 0.0001 的 Adam 优化器训练模型。
- 初始学习率为 0.00005 , 并在 37500 次迭代时降低 0.1 倍。
- $\lambda_{cls} = 2, \lambda_{L1} = 5, \lambda_{giou} = 2.$
- proposal box , proposal feature 和 stage 的数量为 500 , 500 , 6.
- 关系信息提取器的维度为 256。
- $s = 0.7, \theta = 0.4.$



## 在 CorwdHuman 的实验结果

Method	#Queries	AP	$MR^{-2}$	JI
<i>box-based</i>				
RetinaNet [26]	-	85.3	55.1	73.7
ATSS [50]	-	87.0	51.1	75.9
ATSS [50]+MIP [8]	-	88.7	51.6	77.0
FPN [25]+NMS	-	85.8	42.9	79.8
FPN [25]+soft NMS	-	88.2	42.9	79.8
FPN+MIP [8]	-	90.7	41.4	82.4
FPN <sup>†</sup> +NMS	-	84.9	46.3	—
Adaptive NMS <sup>†</sup> [28]	-	84.7	47.7	—
PBN <sup>†</sup> [21]	-	89.3	43.4	—
<i>point-based</i>				
FCOS [42]	-	86.8	54.0	75.7
FCOS [42]+MIP [8]	-	87.3	51.2	77.3
POTO [46]	-	89.1	47.8	79.3

<i>query-based</i>					
DETR [3]	100	75.9	73.2	74.4	
PEDR [24]	1000	91.6	43.7	83.3	
D-DETR [54]	1000	91.5	43.7	83.1	
S-RCNN [39]	500	90.7	44.7	81.4	
S-RCNN [39]	750	91.3	44.8	81.3	
S-RCNN+ <i>Ours</i>	500	92.0	<b>41.4</b>	83.2	
S-RCNN+ <i>Ours</i>	750	<b>92.5</b>	41.6	83.3	
D-DETR+ <i>Ours</i>	1000	92.1	41.5	<b>84.0</b>	

Table 2: Comparisons of different methods on *CrowdHuman* validation set, +MIP represents multiple instance prediction with set NMS as post-processing. <sup>†</sup> indicates the approach is implemented by PBM [21]. S-RCNN – Sparse RCNN [39]. D-DETR – deformable DETR [54].



## 在 CorwdHuman 的实验结果

Method	#Queries	AP	$MR^{-2}$	JI
<i>box-based</i>				
RetinaNet [26]	-	85.3	55.1	73.7
ATSS [50]	-	87.0	51.1	75.9
ATSS [50]+MIP [8]	-	88.7	51.6	77.0
FPN [25]+NMS	-	85.8	42.9	79.8
FPN [25]+soft NMS	-	88.2	42.9	79.8
FPN+MIP [8]	-	90.7	41.4	82.4
FPN <sup>†</sup> +NMS	-	84.9	46.3	—
Adaptive NMS <sup>†</sup> [28]	-	84.7	47.7	—
PBN <sup>†</sup> [21]	-	89.3	43.4	—
<i>point-based</i>				
FCOS [42]	-	86.8	54.0	75.7
FCOS [42]+MIP [8]	-	87.3	51.2	77.3
POTO [46]	-	89.1	47.8	79.3

<i>query-based</i>					
DETR [3]	100	75.9	73.2	74.4	
PEDR [24]	1000	91.6	43.7	83.3	
D-DETR [54]	1000	91.5	43.7	83.1	
S-RCNN [39]	500	90.7	44.7	81.4	
S-RCNN [39]	750	91.3	44.8	81.3	
S-RCNN+ <i>Ours</i>	500	92.0	<b>41.4</b>	83.2	
S-RCNN+ <i>Ours</i>	750	<b>92.5</b>	41.6	83.3	
D-DETR+ <i>Ours</i>	1000	92.1	41.5	<b>84.0</b>	

Table 2: Comparisons of different methods on *CrowdHuman* validation set, +MIP represents multiple instance prediction with set NMS as post-processing. <sup>†</sup> indicates the approach is implemented by PBM [21]. S-RCNN – Sparse RCNN [39]. D-DETR – deformable DETR [54].



## 消融实验

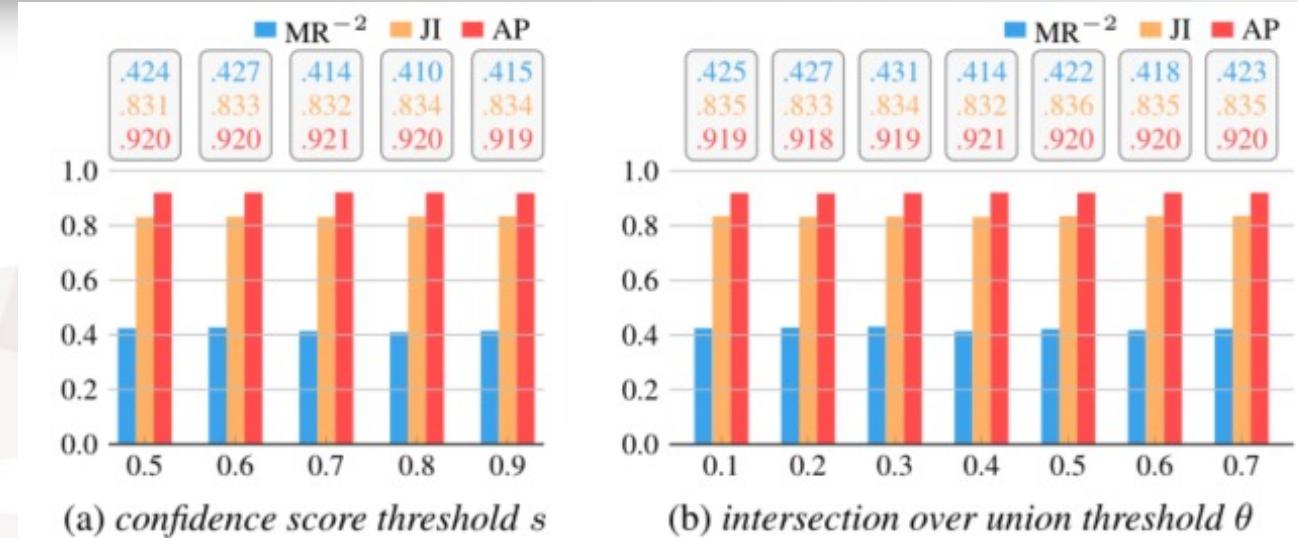


Figure 7: Performance of the proposed method with different configurations of hyper-parameter  $s$  and  $\theta$  on *CrowdHuman* [37] dataset.

$\mathcal{R}$	LMSA	$E$	AP	MR <sup>-2</sup>	JI
			90.7	44.7	81.4
✓			91.5	43.0	83.0
✓	✓		<b>92.0</b>	42.0	<b>83.5</b>
✓	✓	✓	<b>92.0</b>	<b>41.4</b>	83.2

(a) Ablations of different modules.

Method	#Queries	AP	MR <sup>-2</sup>	JI
GossipNet [18]	-	80.4	49.4	81.6
RelationNet [20]	-	81.6	48.2	74.6
IterDet [35]	-	88.0	47.5	78.0
D-DETR+Ours	500	91.2	42.6	<b>84.0</b>
S-RCNN+Ours	500	<b>92.0</b>	<b>41.4</b>	83.2

(b) Comparisons of different relation modeling approaches.

Method	#Queries	MR <sup>-2</sup>	AP
FPN+NMS	-	9.8	94.7
FPN+Soft-NMS [1]	-	9.9	94.9
MIP [8]		8.8	95.8
D-DETR [54]	500	9.4	96.6
S-RCNN [39]	500	10.0	96.8
D-DETR+Ours	500	7.8	96.7
S-RCNN+Ours	500	<b>7.8</b>	<b>97.6</b>

(c) Performance comparisons on *CityPersons*.



4

# References



## 参考

- <https://zhuanlan.zhihu.com/p/482451342>

## 源码

- <https://github.com/megvii-research/Iter-E2EDET>



謝謝

Thank You

THANKS

