A. Implementation of Deformable DETR with progressive predicting method.

We also deploy our progressive predicting approach on the Deformable DETR [10] to demonstrate the generality of our method. Similar to Sparse RCNN [8], the decoder in Deformable DETR consists of 6 stages, whose decoding stages are depicted in Figure. 1a. As described in the manuscript, we integrate our designed components into the last decoding stage. Figure. 1b also describe its detail architecture. For the hyper-parameters setting, .e.g. confidence score threshold s, are identical to those adopted in Sparse RCNN [8].

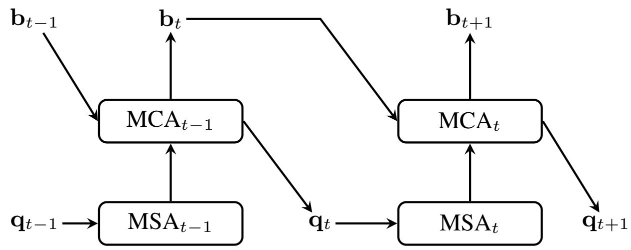
We choose the Deformable DETR with iterative bounding box refinement. Following Deformable DETR [10], we use *ResNet-50* [3] pre-trained on *ImageNet* [2] as backbone. The model is trained with Adam optimizer and a weight decay of 0.0001. The total training duration is 50 epochs on 8 GPUs with 1 image per GPU. The initial learning rate is 0.0002 and dropped by a factor of 0.1 after 40 epochs. The parameters initialization in the newly added components and losses weights are identical to the original work [10]. The default number of queries and stages is 500 and 6, respectively. The hyper-parameters s and θ are also 0.7 and 0.4, respectively. The gradients are detached at proposal boxes from the second stage to stabilize training. We stop gradient back-propagation from the last stage to the previous decoding stages. Besides, those negative samples who overlaps with ignore region with an *intersection-over-area*(IoA) greater than 0.7 are not involved in training.

B. Performance change of a query-based decoder when handling crowded scenes.

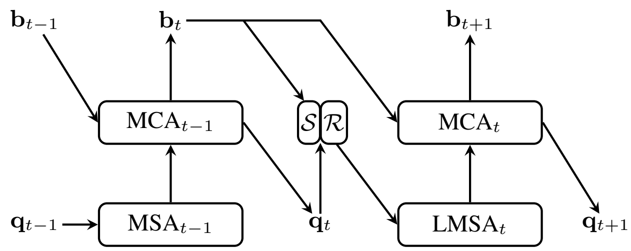
The performance of a *query-based* detector would not be improved but will degrade as the depth of a decoder increases when handling crowded scenes. Experiments are conducted on *CrowdHuman* dataset, taking Sparse RCNN based on *ResNet-50* as base detector. It equips with 500 queries. We adjust the depth of its decoder while keeping the others unchanged. As is described in Table. 1, the performance degrades as the depth of the decoder increases.

#Depth	#Queries	AP	MR^{-2}	JI
6		90.7	44.7	81.4
7		90.6	45.7	81.0
8	500	90.4	45.9	80.3
9		90.7	44.4	80.9
10		90.2	46.6	80.0

Table 1: Experiment analysis as the depth of a decoder increases, which performs on *CrowdHuman* dataset.



(a) Decoder in deformable DETR [10]. MCA – multi-head cross-attention, MSA – multi-head self-attention.



(b) Decoder in SR-Deformable DETR (Ours). $\mathcal{S}$ – Prediction Selector, $\mathcal{R}$ – Relation information extractor, LMSA – local multi-head self-attention.

Figure 1: 1a is the architecture of decoding stage in deformable DETR [10]; 1b describes the decoding stage structure equipped with our designed components for progressive predicting schema.

C. Performance of query detector with large model in crowded scenes.

To explore the detection upper bound of a query-based detector in handling crowded scenes, we replace the *ResNet-50* [3] with a large backbone, Swin-Large [5]. Experiments are conducted on *CrowdHuman* [7] and *CityPersons* [9] datasets, with the same training strategy described in the manuscript. As depicted in Table. 2, our method can significantly boost the performance of a query-based detector equipped, which achieves a *state-of-the-arts* results on both *CrowdHuman* and *CityPersons* validating datasets.

Method	Dataset	#Queries	AP	MR^{-2}	JI
S-RCNN		500	93.1	39.9	85.1
D-DETR	*CHuman*	1000	93.8	**37.4**	86.5
S-RCNN+*Ours*		500	93.4	39.6	86.3
D-DETR+*Ours*		1000	**94.1**	37.7	**87.1**
S-RCNN		500	98.3	5.9	93.7
D-DETR	*CPersons*	500	96.4	8.4	92.0
S-RCNN+*Ours*		500	**98.4**	**4.9**	**94.2**
D-DETR+*Ours*		500	97.5	5.9	93.7

Table 2: Experiment on *CHuman*(*CrowdHuman*) and *CPersons*(*CityPersons*) with Swin-L [5]. S-RCNN – Sparse RCNN [8], D-DETR – Deformable DETR [10]

Figure 2: Results visualization of RelationNet [4], IterDet [6], Sparse RCNN [8], deformable DETR [10] and our approach based on them [8, 10]. Blue boxes are true positive detections, light yellow boxes are missed instances and orange boxes are false positives. Green boxes represent progressively refined detections in our method.

References

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

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[4] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 2

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 1

[6] Danila Rukhovich, Konstantin Sofiiuk, Danil Galeev, Olga Barinova, and Anton Konushin. Iterdet: Iterative scheme for objectdetection in crowded environments. *CoRR*, abs/2005.05708, 2020. 2

[7] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 1

[8] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, and Ping Luo. Sparse r-cnn: End-to-end object detection with learnable proposals. *arXiv preprint arXiv:2011.12450*, 2020. 1, 2

[9] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4457–4465, 2017. 1

[10] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR 2021: The Ninth International Conference on Learning Representations*, 2021. 1, 2