DETR for Pedestrian Detection

Matthieu Lin [1] Chuming Li [2] Xingyuan Bu [2] Ming Sun [2] Chen Lin [3]
Junjie Yan [2] Wanli Ouyang [4] Zhidong Deng [1]
[1]Department of Computer Science and Technology, Tsinghua University.*
[2]SenseTime Group Limited, [3]University of Oxford , [4]The University of Sydney
{lin-yh19@mails, michael@}.tsinghua.edu.cn
{lichuming, sunming1, yanjunjie }@sensetime.com
sefira32@gmail.com, chen.lin@eng.ox.ac.uk, wanli.ouyang@sydney.edu.au

Abstract

Pedestrian detection in crowd scenes poses a challenging problem due to the heuristic defined mapping from anchors to pedestrians and the conflict between NMS and highly overlapped pedestrians. The recently proposed end-to-end detectors(ED), DETR and deformable DETR, replace hand designed components such as NMS and anchors using the transformer architecture, which gets rid of duplicate predictions by computing all pairwise interactions between queries. Inspired by these works, we explore their performance on crowd pedestrian detection. Surprisingly, compared to Faster-RCNN with FPN, the results are opposite to those obtained on COCO. Furthermore, the bipartite match of ED harms the training efficiency due to the large ground truth number in crowd scenes. In this work, we identify the underlying motives driving ED's poor performance and propose a new decoder to address them. Moreover, we design a mechanism to leverage the less occluded visible parts of pedestrian specifically for ED, and achieve further improvements. A faster bipartite match algorithm is also introduced to make ED training on crowd dataset more practical. The proposed detector PED(Pedestrian End-to-end Detector) outperforms both previous EDs and the baseline Faster-RCNN on CityPersons and CrowdHuman. It also achieves comparable performance with state-of-the-art pedestrian detection methods. Code will be released soon.

1. Introduction

Pedestrian detection is a critical research field due to its wide application in self-driving, surveillance and robotics.

Figure 1: The green box is the visible region of the person behind. The blue box and dot are the anchor and point matching the person. As we can observe, the point or point of the anchor is lying on the body of the front person, hence making the mapping ambiguous. While deformable DETR is able to adaptively refine its attention position to the features of the visible part (red point)

In recent years, promising improvements of pedestrian detection have been made. However, pedestrians in occluded or crowd scenes remain difficult to detect accurately.

Pedestrian detection involves two fundamental challenges that remain to be addressed: (1) mapping features to instances and (2) duplicate prediction removal. In the former case (1), most detectors build a heuristic mapping from points on the convolution neural network (CNN) feature maps to ground truth (GT) bounding box. A point is assigned to a GT if it has a small distance to the GT's center or the anchor defined on it has a high intersection over union(IOU) with the GT. Nevertheless, due to the highly overlapped pedestrians with appearance variance. Points lying in the central part of one GT is likely to be mapped to another one. It means the heuristically defined mapping is

*State Key Laboratory of Intelligent Technology and Systems, Beijing National Research Center for Information Science and Technology, and Center for Intelligent Connected Vehicles and Transportation.

ambiguous. In Figure 1, we show that both distance-based mapping and anchor-based mapping suffer from ambiguity even with visible annotations. In the latter case (2), duplicate proposals of a single GT is usually provided by modern detectors and need a post-process mechanism to filter out. However, the widely used non-maximum suppression(NMS) relies on intersection-over-union(IOU) and fails in crowd scenes. Because duplicate proposals and another GT's proposals may both have high IOUs with a GT's true positive proposal.

Existing solutions to pedestrian detection predominantly focus on two types of improvements. Some researchers explore using more distinctive body parts, *e.g.*, head or visible region, and use it to learn extra supervision, re-weight feature maps or guide the anchor selection [40, 28, 5]. Some other works propose clever methods to introduce more signals to make duplicate proposals and close GT's proposals more distinguishable, including neighbor GT's existence and direction, IOU between visible regions and local density [24, 17, 5]. Both the two types of works achieve significant improvements and partly solve the two challenges mentioned above.

The recently proposed end-to-end detectors, DETR and deformable DETR [4, 43], perform comparably or even better on common objects detection. In the subsequent part, we use 'DETR' to refer to DETR and deformable DETR. We identify DETRs' two properties which imply their natural advantages on pedestrian detection over the former works. On one hand, DETR is query-based and does not rely on heuristic design of the mapping between feature map points and GTs. Instead, DETR queries adaptively determine their effective attention areas feature maps and the corresponding objects. On the other hand, DETR learns a bipartite match between queries and GTs. The match assigns an individual query for each GT, without any duplicate proposal. The two properties suggest DETR's potential in solving the two challenges of pedestrian detection specifically.

Model	Epochs	GPU days	AP	MR^{-2}
Faster-RCNN	20	0.75.	85.0	50.4
DETR	300	223.7	66.12	80.62
+Deformable	50	8.4	86.74	53.98

Table 1: Comparison of DETRs and Faster-RCNN + FPN on CrowdHuman. All models are trained on 8 Tesla V-100s. For both DETRs we increase the number of object queries to 400 due to the large GT number of CrowdHuman.

We explore DETR's performance on pedestrian detection and compare them with Faster-RCNN [30], which is the standard baseline used in both DETR works and pedestrian detection works. Unfortunately, the results are opposite of that on COCO [23]. Table 1 indicates that both original DETR and deformable DETR perform much worse than Faster-RCNN on CrowdHuman [31]. Additionally, the training of DETR on dataset of crowd pedestrians is quite time-consuming due to the standard KM algorithm [1] used for bipartite match that has a time complexity cubic of the GT number. The bipartite match costs about 2 times the time of the detectors' forward plus backpropagation on CrowdHuman, thus bottlenecks DETR's training.

We analyse the reasons behind deformable DETR's poor performance on pedestrian detection due to its advantages over original DETR. We find that its (1) sparse uniform queries and (2) weak attention field harm the performance. The decoder of deformable DETR learns a mapping from sparse uniformly distributed queries to a naturally local dense pedestrian cluster. As we discuss in Section 4.2, such mapping is ambiguous and results in missed GTs. The attention positions of the decoder over the feature maps is also problematic. The attention positions is adaptively learnt during training, while they do not converge to a rectified and compact position set which covers the corresponding GT well. On the contrary, they tend to cover more than one GTs or not extensible enough for large persons. From these observations, we propose a decoder with **d**ense **q**ueries and **r**ectified attention **f**ield (DQRF). DQRF significantly improves DETR on pedestrian detection and closes the gap between DETR and Faster-RCNN. Furthermore, we explore how to leverage annotations of visible regions and establish a visible region based set supervision, namely V-Match, together with a data augmentation which is visible region aware, which enhance DETR's performance further. Finally, we design a heuristic improvement of KM algorithm based on the prior that GT tends to match its close proposals in the bipartie matching of ED. The resulted Fast-KM gains up to 10x speed-up, making DETR practical on pedestrian detection tasks.

Our contributions are summarized as:

- We conduct in-depth analysis of DETR for pedestrian detection task and identify the problems when directly applying DETR for pedestrian detection.
- We propose a new decoder for DETR, DQRF, which significantly improves DETR on pedestrian detection and closes the gap between DETR and Faster-RCNN. A Fast-KM is also proposed to make DETR practical on pedestrian detection.
- We further explore the leverage of annotations of visible regions specifically for DETR and establish a visible region based set supervision, V-Match, together with a data augmentation which is aware of visible region.

The resulted **PE**destrian-specific **D**ETR, namely PED, outperforms competitive Faster-RCNN and achieves compara-

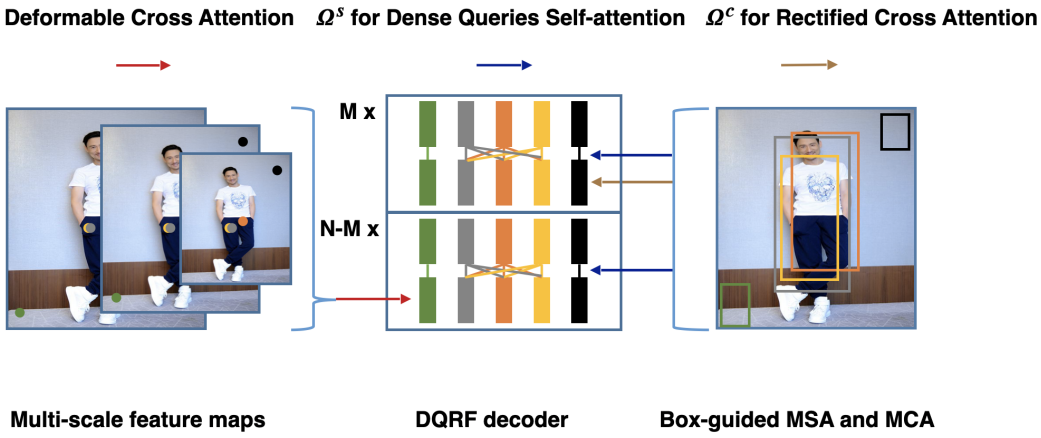

Figure 2: Overview of our DQRF decoder with N DQ layers and M RF layers. The encoder architecture is kept as in [43].

ble performance with state-of-the-art results on the challenging CrowdHuman and CityPersons [39] benchmarks. We hope that our work can serve as a new baseline for end-to-end detectors on pedestrian detection in the same way DETR and deformable DETR for common object detection.

2. Related Work

Generic Object Detection. In the era of deep learning, most generic object detection could be roughly divided into two categories, *i.e.*, two-stage detection [12, 14, 11, 30] and one-stage detection [25, 29, 22, 20], depending on whether an explicit region proposal and pooling process are employed. To further improve the performance, FPN [21, 10] and DCN series [7, 42] are introduced to enhance the feature representation. Meanwhile, iterative prediction [3] and extra supervision [13, 19, 18] could yield more precise bounding boxes. Recently, some works try to remove the predefined anchor hypothesis over the feature map grid, known as anchor-free detection methods. They tend to use the center [9] or keypoint [41, 34] instead of anchor box. Relation networks [16] models the relation between different proposals by long-range attention, which is then used to distinguish whether a generated proposal is the unique prediction. However, it still involves hand-crafted rank and box embedding. Different from the above generic object detection methods, DETR [4] can generate set prediction direct from the input image.

Pedestrian Detection. Pedestrian detection is the fundamental technology for self-driving, surveillance, and robotics. Although great progress has been made in the past decade [8, 38, 39], pedestrian detection in crowd scenario remains challenging due to the confusing features and dilemmatic NMS [31]. OR-CNN [40] and MGAN [28] regard the invisible parts as noise and down-weighting those features to deal with the confusing feature. PedHunter [6] applies a stricter overlap strategy to reduce the ambiguity of matching. Using the human head as a clue also be explored in [6, 5]. For the dilemmatic NMS, adaptive NMS [24] first predicts the crowd density, then dynamically adjusts the NMS threshold according to predicted density. The visible box and head box [17, 5] also increase the performance of the NMS and partly address the essential problem of how to generate an appropriate prediction. Anchor-free methods are also adapted to pedestrian detection in [27, 37], and they reserves NMS. [33] views proposal prediction as a sequence generation, which is an unnecessary attribute for detection.

DETR. Generic object detection methods usually contain a hand-crafted post-process, *i.e.*, NMS. Despite its variant type [2, 15], this post-process has no access to the image information and the network feature and must be employed solely. Thus, NMS can not be optimized in an end-to-end manner. The recent DETR [4] and deformable DETR [43] utilize the encoder-decoder architecture based on the transformer module, which could essentially build context features and remove duplicates. Although they achieve relatively high performance in common detection datasets, they do not work well in pedestrian detection under crowd scenarios, as discussed in the introduction.

3. Revisit DETR

3.1. Decoder

The recent DETRs are based on the architecture of transformer [35] and assign a unique query for each GT through bipartite matching. Here we briefly formulate the detection process of DETR as follows.

Let $\boldsymbol{q}$ represents a set of queries with $\boldsymbol{q}_i \in R^C$ and $\boldsymbol{x}$ represents the feature map points with $\boldsymbol{x}_i \in R^C$. In the decoder

of the transformer, the query set $\boldsymbol{q}$ is iteratively updated via cross-attention between $\boldsymbol{q}$ and $\boldsymbol{x}$ and self-attention among $\boldsymbol{q}$. This process can be formulated as a sequence of functions F^t, where $\boldsymbol{q}^t = F^t(\boldsymbol{q}^{t-1}, \boldsymbol{x})$, and $t \in 1, ..., T$ with T denoting the number of decoder layers. We further decompose F^t into F_c^t and F_s^t. $F_c^t(\boldsymbol{q}, \boldsymbol{x})$ represents the cross attention between $\boldsymbol{q}$ and $\boldsymbol{x}$ and $F_s^t(\boldsymbol{q})$ represents the self attention among $\boldsymbol{q}$. Hence, $F^t(\boldsymbol{q}^{t-1}, \boldsymbol{x}) = F_c^t(F_s^t(\boldsymbol{q}^{t-1}), \boldsymbol{x})$.

3.2. Multi-Head Attention

The functions F_c^t and F_s^t of DETR are based on transformer's multi-head attention. In both the two DETRs, F_s^t consists of a standard multi-head self attention module(MSA) followed by a multiple linear projection(MLP) as in Eq. 1, 2, where LN means layer normalization.

$$\boldsymbol{q}_{msa}^{t-1} = LN(MSA(\boldsymbol{q}^{t-1}) + \boldsymbol{q}^{t-1}), \quad (1)$$

$$F_s^t(\boldsymbol{q}^{t-1}) = \boldsymbol{q}_s^{t-1} = LN(MLP(\boldsymbol{q}_{msa}^{t-1}) + \boldsymbol{q}_{msa}^{t-1}). \quad (2)$$

The cross attention function F_c^t similarly consists of a multi-head cross attention module(MCA) between $\boldsymbol{q}$ and $\boldsymbol{x}$ followed by a MLP (Eq. 3, 4).

$$\boldsymbol{q}_{mca}^{t-1} = LN(MCA(\boldsymbol{q}_s^{t-1}, \boldsymbol{x}) + \boldsymbol{q}_s^{t-1}), \quad (3)$$

$$F_c^t(\boldsymbol{q}^{t-1}, \boldsymbol{x}) = \boldsymbol{q}^t = LN(MLP(\boldsymbol{q}_{mca}^{t-1}) + \boldsymbol{q}_{mca}^{t-1}). \quad (4)$$

Both MSA and MCA can be expressed by a basic multi-head attention(MA) module as in Eq. 5, 6, 7.

$$MA(\boldsymbol{q}_i, \boldsymbol{z}) = \sum_{m=1}^{M} \boldsymbol{W}_m \sum_{k \in \Omega_i} \boldsymbol{A}_{mik} \boldsymbol{W}_m^V \boldsymbol{z}_k, \quad (5)$$

$$MSA(\boldsymbol{q}_i) = MA(\boldsymbol{q}_i, \boldsymbol{q}), \quad (6)$$

$$MCA(\boldsymbol{q}_i, \boldsymbol{x}) = MA(\boldsymbol{q}_i, \boldsymbol{x}). \quad (7)$$

$\boldsymbol{z}$ is a set of vectors where $\boldsymbol{q}_i$ will make attention on. In MSA, $\boldsymbol{z}$ is $\boldsymbol{q}$ itself and in MCA the feature maps $\boldsymbol{x}$. For each head m, a linear projection $\boldsymbol{W}_m^V \in R^{\frac{C}{M} \times C}$ is operated on $\boldsymbol{z}$ to map it to a new representation. An attention weight $\boldsymbol{A}_{mik}$ is applied to weighted sum the representations on positions in Ω_i, which is the attention field of $\boldsymbol{q}_i$. Finally, the weighted summed features of all heads are linearly projected to R^C via $\boldsymbol{W}_m$ and summed.

Original DETR and deformable DETR differs in the designs of MCA, where the MA modules have different formulations of Ω_i and $\boldsymbol{A}_{mik}$. In deformable DETR, Ω_i is a small set of fractional positions and it cover points on multiple feature maps with different resolutions, and Ω_i is obtained via linear projection over the query $\boldsymbol{q}_i$. However, in DETR Ω_i is all points on the feature map with the lowest resolution. The former way is more efficient and hence supports high resolutions. Moreover, $\boldsymbol{A}_{mik}$ of deformable DETR is also obtained via linear projection, and has a lower computaion cost than DETR, which projects $\boldsymbol{q}_i$ and $\boldsymbol{z}_k$ into new representations and calculates the inner product between them.

3.3. Set Prediction

In DETR, the query set $\boldsymbol{q}^t$ at each decoder layer is projected into a bounding box set b^t, via two MLPs separately for classification and box regression. Each box set b^t is supervised via the bipartie matching between b^t and GTs. For the bipartie matching, both DETR and Deformable DETR use standard KM algorithm.

4. Method

4.1. Why DETR Fails in Pedestrian Detection?

We compare both original DETR and deformable DETR on CrowdHuman, a challenging dataset containing highly overlapped and crowd pedestrians, and compare them with Faster-RCNN, a standard baseline in pedestrian detection. All the detector are implemented following the standard hyper-parameters. A counter-intuitive result is illustrated in Table 1. Both original DETR and deformable DETR have obvious performance drop compared with their baselines. Moreover, the training efficiency of DETR on CrowdHuman is much lower than on the common object detection dataset, COCO.

As deformable DETR shows higher performance than original DETR, we take it as our starting point and analyse the reasons behind its performance drop. Our investigation focuses on the decoder, which is the key architecture of DETR's set prediction mechanism. We define $\Omega_i^{c,t}$ as the cross attention field of the i-th query on the feature map and $\Omega_i^{s,t}$ as the positions set the i-th query makes self attention on in the t-th decoder layer. In the decoder, the query set $\boldsymbol{q}^0$, $\boldsymbol{q}^1$, ..., $\boldsymbol{q}^T$ is iteratively updated. In each decoder layer t, $\boldsymbol{q}^{t-1}$ exchange information among each other via self attention over $\Omega^{s,t-1}$, then predict the attention field $\Omega^{c,t-1}$ to extract information of objects from feature maps. At last layer T, each query $\boldsymbol{q}_i^T$ matches a GT or background with its attention trajectory $\Omega_i^{c,0}$, $\Omega_i^{c,1}$, ..., $\Omega_i^{c,T-1}$.

4.2. Dense Queries

First, we find a conflict between the locally **dense** distribution of GTs and uniformly **sparse** distribution of query set $\boldsymbol{q}$. Specifically, after training, the initial cross attention fields $\Omega^{c,0}$ of different queries are uniformly distributed on the feature map as shown in Figure 3(left), due to the uniform appearance of pedestrians. The queries are also sparse due to the limit of computation resources. However, in a single image, pedestrians tend to be distributed densely in some local regions(e.g. a corner or a horizontal line) naturally, while the number of queries $\boldsymbol{q}^0$ whose

attention field $\Omega^{c,0}$ locates initially in such local dense regions is not always enough to match all pedestrians lying in them, as in Figure 3. It means, the decoder layers F^t learns to shrink the attention fields $\Omega^{c,t-1}$ of the uniformly distributed sparse queries progressively from the whole image to compact dense object clusters. This process has two requirements: (1) a vast perception range of objects and (2) a mapping from a vast range of initial positions to local dense GTs. Two conflicts stands in the two requirements. On one hand, the vast perception range suggest strict requirement on the CNN's reception field. On the other hand, the mapping is highly ambiguous because there are few prior geometry cues for queries to decide how GTs are assigned among them, indeed, although queries are supervised by strict bipartie matching, there is still imbalance between the number of queries lying on different GTs in the dense object clusters. Some GTs are missed while some others contain more than 3 queries, as observed in Figure 4.

The discussions above imply that dense queries will help. When queries are dense enough, the bipartite matching result of each GT is roughly its nearest unique query and the requirement of the query's reception field is much lower. However, the time complexity of the MSA module is quadratic of the query number and hardly bears a dense query setting. We design a **D**ense **Q**uery(DQ) module to support a dense setting via reducing the complexity of MSA from $O(N_q^2)$ into $O(N_q)$ of the query number N_q.

Local self attention is developed in transformer as an effective way to improve computation efficiency. However, in transformer, queries have one-to-one correspondence to the token positions, and locality can be naturally designed by restricting Ω_i^s as some near positions $\{..., i-1, i, i+1, ...\}$. However, though queries in DETR is equipped with attention positions on feature map, the positions are fractional and variable during training. To develop a distance measure for queries in DETR, we first review what queries should be in Ω_i^s for a certain query $\boldsymbol{q}_i$. In MSA, $\boldsymbol{q}_i$ receives information from queries in Ω_i^s to determine whether itself or another query in Ω_i^s is matched to a GT, as discussed in DETR. This reasonable assumption suggests that the distance should be measured by the possibility that two queries will match the same GT. Consider that each decoder layer predicts its box set $\boldsymbol{b}_i^t$ sequentially, we use the overlaps between $\boldsymbol{q}_i^{t-1}$'s box prediction $\boldsymbol{b}_i^{t-1}$ and $\boldsymbol{q}_j^{t-1}$'s $\boldsymbol{b}_j^{t-1}$ in the former decoder layer, to measure the distance of $\boldsymbol{q}_i^{t-1}$ and $\boldsymbol{q}_j^{t-1}$, because the higher the overlap the more possible that $\boldsymbol{q}_i^{t-1}$ and $\boldsymbol{q}_j^{t-1}$ predict the same GT. Hence we define the distance measure d_{ij}^{t-1} and $\Omega_i^{s,t-1}$ as:

$$d_{ij}^{t-1} = 1 - GIOU(\boldsymbol{b}_i^{t-1}, \boldsymbol{b}_j^{t-1}), \tag{8}$$

$$\Omega_i^{s,t-1} = \{\boldsymbol{\tau}_{i1}^{t-1}, \boldsymbol{\tau}_{i1}^{t-1}, ..., \boldsymbol{\tau}_{iK}^{t-1}\}, \tag{9}$$

where $\boldsymbol{\tau}_i^{t-1}$ is the ascending order of $\boldsymbol{d}_i^{t-1}$ and we select the nearest K neighbors of $\boldsymbol{q}_i^{t-1}$ based on the defined $\boldsymbol{d}_i^{t-1}$. As shown in Section 5, our DQ algorithm supports twice more queries without calculation cost increase, and achieves even better performance than simply adding twice more queries without changing Ω_i^s via forcing queries to focus only on nearby queries.

4.3. Rectified Attention Field

Another problem arises in DETR is that the cross attention $\Omega_i^{c,T}$. $\Omega_i^{c,t}$ is predicted via linear projection over the query feature, it bears a risk to be messy or narrow. In our experiment, in average 34.9% attention positions in $\Omega_i^{c,t}$ is out of the box of the GT its corresponding query $\boldsymbol{q}_i^t$ matched and 69.7% of them lies on another nearby GT's box as in Figure 5(left), which introduces noise. Furthermore, the learned $\Omega_i^{c,t}$ is often not wide enough for the queries which match large peoples, as in Figure 5(right), it harms the accuracy of both the classification score and box regression.

To relieve the noisy or narrow attention field of the queries, we design a RF(**R**ectified attention **F**ield) module to rectify the attention field $\Omega_i^{c,t}$ of the final M layers. As observed in Table 2, we find that more than 95% queries match the same GTs at the last three layers. It means the box prediction $\boldsymbol{b}_i^t$ at the 4-th or 5-th layer is nearly always around its final target GT, and we can use the intermediate box prediction $\boldsymbol{b}_i^t$ to get a more compact while wide enough attention field $\Omega_i^{c,t}$. We set the attention field $\Omega_i^{c,t}$ as:

$$\Omega_i^{c,t-1} = \{(x_i^{t-1} + \frac{i1}{R+1} w_i^{t-1}, y_i^{t-1} + \frac{i2}{R+1} h_i^{t-1})\}, \\ i1, i2 \in \{1, ..., R\}, \tag{10}$$

where x_i^{t-1}, y_i^{t-1}, w_i^{t-1}, h_i^{t-1} are position, width and height of the box prediction $\boldsymbol{b}_i^{t-1}$. We use uniform distibuted $R \times R$ points among $\boldsymbol{b}_i^{t-1}$, it relieves the risk of a learned $\Omega_i^{c,t-1}$ to be messy or narrow. Table 4 shows that our proposed method improves significantly over the baseline on CrowdHuman. The proposed decoder, namely DQRF, is shown in Figure 2.

	2	3	4	5	6
Similarity	0.939	0.948	0.957	0.964	0.970
IoU	0.513	0.629	0.754	0.841	0.908

Table 2: Similarity ratio of matched ground truth between each layer and their previous layer and IoU overlap between each predicted box with their previous layer. Values are computed via averaging over the train set with a pre-trained Deformable DETR on CrowdHuman.

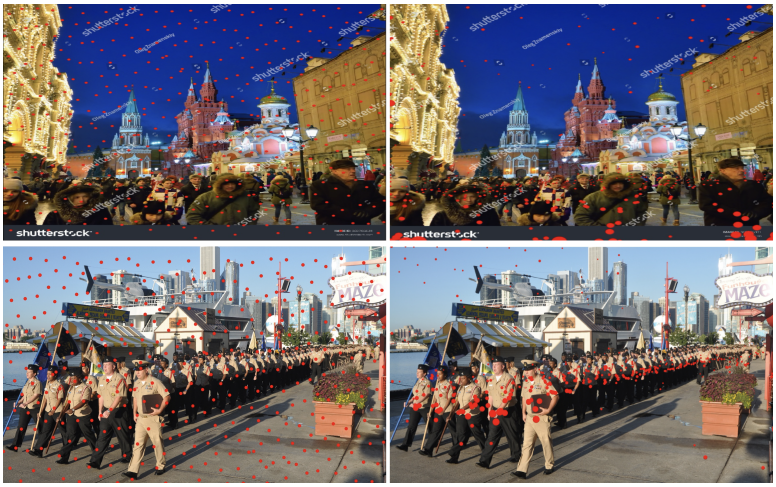

Figure 3: Visual comparisons between the central positions of the attention fileds Ω^c of the queries in the first decoder layer(left) and the last decoder layer(right). We can observe that object queries need to learn a shrinking from sparsely uniform distribution to a dense cluster of pedestrian. The size of each circle is representative of the predicted area of each box predicted by the corresponding query.

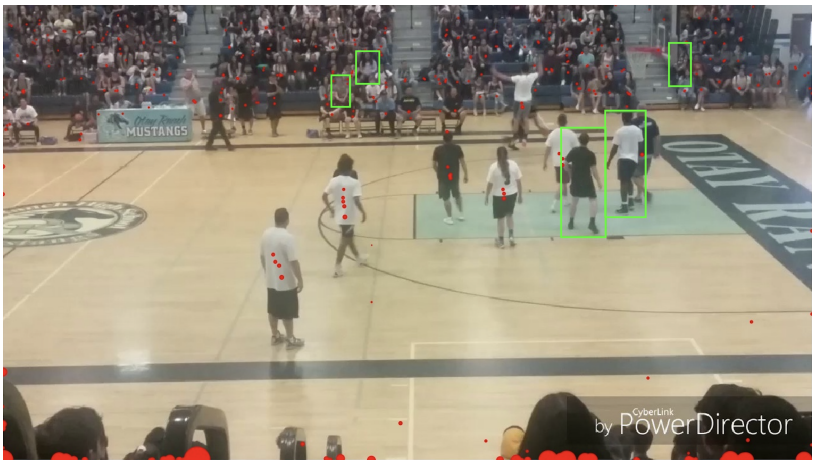

Figure 4: Prediction of the baseline deformable DETR. We highlight in green GTs that are missed while many other GTs contain more than 3 queries.

4.4. Further improvement with visible region

Recent works [17, 5] develop methods to leverage visible region annotations. These methods reveal that, with marginal extra cost, rational utilization of visible annotations leads to considerable gains. As such, we also propose to leverage visible annotations under the end-to-end framework. The proposed method, **V-match**, achieves similar performance gain and introduces no extra computational cost.

We design a novel adaptation of the targets of DETR. Considering DETR predicts a sequential sets of boxes $\boldsymbol{b}_1, ..., \boldsymbol{b}_T$, we assign the supervision of full boxes on the last L layers while visible boxes on the first $T-L$ layers. It means in the first $T-L$ layers the regression heads predicts visible boxes and the queries are explicitly constrained to focus on the visible part of pedestrians. As shown in Section 5.1, V-match achieves stable improvements at zero cost.

4.5. Other Adaptations

The training of DETR on dataset with crowd pedestrians has low GPU utilization, because the standard KM algorithm for bipartite matching is cubic of the number of GT. We use the prior that a GT is assigned to the predicted box

Figure 5: Visual comparisons between the baseline and the Rectified Attention. The size of each points represent the weights of each sampled point. Pictures with green boxes show attention predicted by deformable DETR for $k = 4$.

among its nearest predicted boxes and adapt it accordingly, the resulted Fast-KM has up to 10 times speed up compared to standard KM on CrowdHuman. We omit the details as the complexity of KM algorithm and release it in our code.

We keep the crop augmentation used in DETR and deformable DETR. While it helps little for traditional detectors, it provides diversity of the objects' distribution to prevent DETR from over fitting the objects' distribution in the training dataset. Nevertheless, it is harmful in crowd pedestrian detection since the occluded pedestrian tends to have a small visible part, which is easy to be cropped out. Thus, we adapt the crop operator to conserve at least 80% area of the visible part of each pedestrian.

5. Experiments

Datasets. We evaluate our PED on two human detection benchmarks: CrowdHuman and CityPersons. These two datasets both contains these two categories of bounding boxes annotations: human visible-region and human full-body bounding boxes. As shown in table 3 Compared to CityPersons, The CrowdHuman dataset is much more challenging as it contains more instances per images, and those instances are often highly overlapped.

Dataset	#person/img	#overlaps/img
CrowdHuman	22.64	2.40
CityPersons	6.47	0.32

Table 3: Statistics of each dataset. The threshold for overlap is IoU > 0.5

Evaluation Metrics. We evaluate the performance of our PED using two standard metrics used for Pedestrian detection, *e.g.*:

- Average Precision following the standard COCO evaluation metric, which computes the area under the curve of the interpolated precision w.r.t recall curve. This metric is the most commonly used metric in metric in detection as it reflects both the precision and the recall.
- Log-average miss rate (MR^{-2}), which computes on a log-scale the miss rate on false positive per image with a range of $[10^{-2}, 10^{0}]$. This metric is the most commonly used metric in pedestrian detection as it reflects the amount of pedestrian that are not detected.

Detailed Settings. Due to the required extra-long training schedule of DETR, we chose to experiment our proposed methods based on Deformable DETR with Iterative Bounding Box Refinement. For all ablation studies, we use as baseline the standard Deformable DETR with Iterative Bounding Box Refinement, all hyper-parameter settings follow Deformable DETR except that we increase the number of queries from 300 to 400, and set the number of queries to 1000 for experiments on our Dense Query method mentioned in Section 4.2. For training, we use the same protocol as in Deformable DETR, *e.g.*, models are trained for 50 epochs with a learning rate drop by a factor of 10 at the 40^{th} epoch. We also slightly modify the original cropping such that full-body bounding-box are not cropped, since they frequently exceed the size of the images. For our final result, since DETR benefits from longer training schedule we also propose to train the model for 100 epoch with a learning rate drop by a factor of 10 at the 90^{th} epoch. We use our DQ setting, set L to 2 for full box supervision and use 3 RF layers.

5.1. Ablation Study

We perform ablation studies and report highest accuracy during training for our new proposed methods in Section 4 on the CrowdHuman dataset. For ablation studies, we replace layers of the deformable DETR with iterative bounding box refinement with our proposed methods starting from the last layer as it is supposed to be the most accu-

rate bounding box prediction. As shown in Table 2, at the latter layers, the predicted bounding boxes are less likely to fluctuate. We hypothesize that replacing first layers does not improve further as table 2 shows that bounding boxes at the first layers are very noisy and the variation (IoU is low) is high.

Ablation study on Rectified Attention Field. As discussed in 4.3 the deformable mechanism might yield noisy attention positions among more than one persons and can make the instance inside a detected box ambiguous, hence we propose to add our Rectified Attention Field to adapt decoder layers. Table 4 shows ablations of our new introduced method by varying the number of decoder layers with our added Rectified Attention Field, our method improves significantly on the baseline under all settings of RF.

#RF layer	0	1	2	3	5
AP	86.74	87.70	87.87	**88.20**	87.76
MR^{-2}	53.98	49.62	48.05	**46.99**	47.38

Table 4: Effect of the number of Rectified Attention Field(RF) layers. Note that the first row is equivalent to the baseline

Ablation study on V-Match. Table 5 shows ablations for the proposed V-match in Sec 4.4 for different values of L. We can observe that for any setup introducing supervision with visible box improves on the baseline.

L	6	5	4	3	2	1
AP	86.74	87.47	87.38	87.35	**87.67**	87.49
MR^{-2}	53.98	49.78	49.25	48.94	**48.07**	48.54

Table 5: Effect of the number of layers with full box supervision. Note that in the case L=6, it is equivalent to the baseline since our baseline has 6 decoder layers

Ablation study on Dense Queries. Table 6 shows ablation for the Dense Queries algorithm. As mentioned, for fair comparison, we train another baseline Deformable DETR with iterative bounding box refinement and increase the number of queries from 400 to 1000. Our baseline with 1000 queries significantly improves on the baseline with 400 queries, supporting our assumption that a dense query approach is needed. We vary the number of layers with DQ to reduce the computational cost, and observe that applying 5 DQ layers on the baseline with 1000 queries even improves further as it forces queries to only attention on nearby queries rather than irrelevant ones or background. K is fixed to 100 after cross validation.

Ablation studies on Crop Augmentation. As discussed in 4.5 we keep at least 80% of the visible area of each pedestrian as occluded pedestrian displays only a small visible part, which is easily cropped out Table 7 shows ablations for our new proposed data augmentation.

# DQ layers	0(no DQ)	0	1	3	5
AP	86.74	88.34	88.81	**89.02**	88.93
MR^{-2}	53.98	49.38	48.37	**47.29**	47.54

Table 6: Effect of the number of layers with our Dense Queries method. Note the first column is equivalent to the baseline with 400 queries and the latter columns are applying different number of layers of DQ on the 1000 queries baseline.

Methods	AP	MR^{-2}
Baseline	86.74	53.98
with Crop Augmentation	**87.17**	**51.95**

Table 7: Ablation study on our Crop Augmentation.

5.2. Results on CityPersons and CrowdHuman

CityPersons contains 2,975 and 500 images for training and validation respectively, and CrowdHuman contains 15,000 and 4370 images for training and validation respectively.

Comparisons with the standard Faster-RCNN+FPN baseline along with other state-of-the-art methods on CrowdHuman & CityPersons. We report results on the Heavy occluded subsets of CityPersons while on CrowdHuman we report AP and MR^{-2}. Table 8 and 9 show our final results on CrowdHuman and CityPersons. Our proposed PED improves by a large a margin on pedestrian Detection compared with deformable DETR and Faster R-CNN, and is even able to compete with very competitive state-of-the-art methods such as PBM, while having comparable FLOPs.

Method	OR-CNN [40]	TLL [32]	RepLoss [36]
Heavy	55.7	53.6	56.9
Methods	ALFNet [26]	CSP [27]	Ours
Heavy	51.9	49.3	**47.70**

Table 8: Results on Heavy subset on CityPerson shows that our method is effective in occluded and crowded scene.

6. Conclusion

In this paper, we design a new decoder, DQRF, which can be easily implemented and helps alleviate the identified drawbacks of DETR on pedestrian detecion. We also propose a faster bipartite matching algorithm and leverage visible box annotations specifically for DETR to get further improvements. We hope that the resulted detector, namely

Method	AP	MR^{-2}	Recall
PBM [17]	89.3	**43.3**	93.33
Ours	**90.08**	44.37	**93.95**
Faster-RCNN* [30]	85.0	50.4	90.24
AdaptiveNMS* [24]	84.7	49.7	91.27
Deformable DETR* [43]	86.74	53.98	92.51
Ours*	**89.54**	**45.57**	**94.00**

Table 9: Results on CrowdHuman. * stands for no usage of visible boxes

PED, inspires future work and serves as a baseline for end-to-end pedestrian detection.

References

[1] https://en.wikipedia.org/wiki/Hungarian_algorithm/. 2

[2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017. 3

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: High quality object detection and instance segmentation. *arXiv preprint arXiv:1906.09756*, 2019. 3

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *arXiv preprint arXiv:2005.12872*, 2020. 2, 3

[5] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Relational learning for joint head and human detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10647–10654, 2020. 2, 3, 6

[6] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, Xudong Zou, et al. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In *AAAI*, pages 10639–10646, 2020. 3

[7] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 3

[8] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011. 3

[9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 3

[10] Golnaz Ghiasi, Tsung-Yi Lin, and Quoc V Le. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7036–7045, 2019. 3

[11] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448, 2015. 3

[12] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014. 3

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017. 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015. 3

[15] Yihui He, Xiangyu Zhang, Marios Savvides, and Kris Kitani. Softer-nms: Rethinking bounding box regression for accurate object detection. *arXiv preprint arXiv:1809.08545*, 2, 2018. 3

[16] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3588–3597, 2018. 3

[17] Xin Huang, Zheng Ge, Zequn Jie, and Osamu Yoshie. Nms by representative region: Towards crowded pedestrian detection by proposal pairing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10750–10759, 2020. 2, 3, 6, 9

[18] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6409–6418, 2019. 3

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

[20] Buyu Li, Yu Liu, and Xiaogang Wang. Gradient harmonized single-stage detector. *arXiv preprint arXiv:1811.05181*, 2018. 3

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. 3

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 3

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[24] Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6459–6468, 2019. 2, 3, 9

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

[26] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634, 2018. 8

[27] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2019. 3, 8

[28] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4967–4975, 2019. 2, 3

[29] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 3

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 3, 9

[31] 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. 2, 3

[32] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 536–551, 2018. 8

[33] Russell Stewart, Mykhaylo Andriluka, and Andrew Y Ng. End-to-end people detection in crowded scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2325–2333, 2016. 3

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

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[36] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018. 8

[37] Zixuan Xu, Banghuai Li, Ye Yuan, and Anhong Dang. Beta r-cnn: Looking into pedestrian detection from another perspective. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[38] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *Proceedings of the iEEE conference on computer vision and pattern recognition*, pages 1259–1267, 2016. 3

[39] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017. 3

[40] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653, 2018. 2, 3, 8

[41] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3

[42] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9308–9316, 2019. 3

[43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3, 9