Favoring One Among Equals - Not a Good Idea: Many-to-one Matching for Robust Transformer based Pedestrian Detection

K.N Ajay Shastry[1] K. Ravi Sri Teja[1] Aditya Nigam[2] Chetan Arora[1]
[1]Indian Institute of Technology, Delhi
[2]Indian Institute of Technology, Mandi

Abstract

We investigate the reasons for lower performance of transformer based pedestrian detection models compared to convolutional neural network (CNN) based ones. CNN models generate dense pedestrian proposals, refine each proposal individually, and follow it up with non-maximal-suppression (NMS) to generate sparse predictions. In contrast, transformer models select one proposal per ground-truth (GT) pedestrian box and backpropagate positive gradient from them. All other proposals, many of them highly similar to the selected ones, are passed negative gradient. Though this leads to sparse predictions, obviating the need of NMS, the arbitrary selection of one among many similar proposals, hinders effective training, and lower accuracy of pedestrian detection. To mitigate the problem, instead of commonly used Kuhn-Munkres matching algorithm, we propose Min-cost-flow based formulation, and incorporate constraints such as, each ground truth box is matched to atleast one proposal, and many equally good proposals can be matched to a single ground truth box. We propose first transformer based pedestrian detection model incorporating our matching algorithm. Extensive experiments reveal that our approach achieves a miss rate (lower is better) of 3.7 / 17.4 / 21.8 / 8.3 / 2.0 *on Eurocity / TJU-traffic / TJU-campus / Cityperson / Caltech datasets compared to* 4.7 / 18.7 / 24.8 / 8.5 / 3.1 *by the current SOTA. Code is available at* https://ajayshastry08.github.io/flow_matcher

1. Introduction

Pedestrian detection is one of the first steps in many computer vision problems, viz, autonomous driving, and surveillance. Similar to many other computer vision problems, recent advances in deep neural network architectures, and availability of large training datasets, have led to significant improvement in the performance of pedestrian detection techniques as well [13, 16, 17, 25, 26]. Taking cue from their success in natural language processing tasks [8, 29], researchers have also explored models based on attention-based transformer architectures [36]. The transformer models have been shown to be more robust, with better generalization compared to CNN models for many computer vision tasks, including object detection [4, 24, 39, 44]. However, strangely enough, even though pedestrian detection can be considered as a specialised object detection problem, the accuracy of transformer models is inferior to the CNN models on common benchmark datasets [13, 17]. The focus of this paper is on investigating the reasons for this variance.

(a) (b) (c) (d)

Figure 1. Comparison between matches performed by different many-to-one matching strategies. Fig. 1a shows proposals around a ground truth. Fig. 1b shows the matching obtained by the repetition of ground truths while performing the Hungarian matching [15]. Fig. 1c shows the matching by IoU-based matching strategy, where a fixed IoU value is used while matching. Fig. 1d shows the matching by our proposed approach. The green, blue, and yellow boxes represent the ground truth, proposals, and matches respectively.

CNN based models for object detection in general, as-well-as pedestrian detection specifically, typically output multiple predictions per ground truth. A post-processing step called non maximal suppression (NMS) scans these predicted boxes in the order of their confidence, and deletes any overlapping boxes with lower confidence scores. However, the step can erroneously delete partially occluded pedestrians as well. On the other hand, transformer mod-

els typically contain an encoder and decoder block, where encoder block also outputs proposal boxes along with their confidence. *Top-n* (*n* is a hyper-parameter here) proposals based on confidence scores are used to initialize queries for the decoder. Removal of overlapping boxes now happens through learnt query interactions in various decoder layers. In CNN models, all proposals with more than a certain intersection-over-union (IoU) score with the GT are given positive gradient. However, in transformers, one chooses a single proposal per GT based on the Kuhn-Munkres (also called Hungarian) matching algorithm and backpropagate positive gradient during training. Our investigation reveals that this adhoc one-to-one matching in the proposal generation stage is the source of reduced performance of transformer models on the pedestrian detection task.

In two-stage models like DINO [39], the proposal generation step uses a Region Proposal Network (RPN) that predicts several overlapping boxes around each pedestrian. Choosing one proposal arbitrarily, and sending positive gradient, whereas sending negative gradient from many similar proposals confuses the proposal generation module, leading to inferior proposals. Recall that proposals are used to initialize queries in the decoder. We observe that in the decoder layers there is seldom any significant change from the initialized box coordinates. Hence, bad quality proposals directly lead to inferior predictions, and missed detections.

Based on our analysis, we propose a novel transformer based pedestrian detection model with improved proposal generation. Specifically, we suggest a new matching technique which allows for many-to-one matching between predicted proposal boxes and a GT box. We formulate the matching problem such that each GT box is matched atleast with one proposal box but can be matched with upto k proposal boxes (k is a hyper-parameter). Fig. 1 shows the comparison between multiple many-to-one matching strategies.

Contributions. Our specific contributions are as follows:

1. We investigate inferior performance of transformer models for pedestrian detection, and identify one-to-one matching between GT and proposals as the cause.
2. Based on our analysis, we propose a min cost flow based matching algorithm to allow many-to-one matching between proposals and GT. The algorithm can be used as an alternative to currently used Hungarian algorithm which gives one-to-one matching.
3. We propose a novel transformer based pedestrian detection models incorporating many-to-one matching. In opposition to prior studies claiming inferior performance of transformer models, our transformer model outperforms all CNN based SOTA models on Eurocity / TJU-traffic / TJU-campus / Cityperson / Caltech datasets giving a miss rate (lower is better) of 3.7 / 17.4 / 21.8 / 8.3 / 2.0 compared to 4.7 / 18.7 / 24.8 / 8.5 / 3.1 by the current SOTA CNN model.

2. Related Works

Classical pedestrian detection. The task can be seen as a subset of the generic object detection problem, consisting of the detection and localization of a pedestrian in an image. Classical techniques posed it as an image classification problem using sliding window across the image, and performing classification at each location. Dalal *et al*. [7] extracted the structure of a pedestrian using HOG features and used an SVM for classification. Felzenszwalb *et al*. [10] went a step further by utilizing the information of each body part to conduct detection, with detection dependent on the detection of each body part and its relative orientation.

Modern pedestrian detection techniques. The techniques can be divided into two categories: Single-stage and Two-stage architectures. Later methods [6, 13, 23, 31, 34] utilize a region proposal network (RPN) to generate an initial set of proposal boxes, which are further refined using another classification and regression network. On the other hand, single-stage detectors directly detect pedestrians using the extracted raw features from the backbone. Both styles can be further sub-divided into anchor-based or non-anchor-based techniques. The former methods [25,30] use a predefined set of boxes with varying sizes and aspect ratios to detect the presence of a pedestrian, whereas non-anchor methods [16, 17, 26, 38] use variations of box center and scale to predict box for a pedestrian. CNN-based models have shown better performance over transformers [14, 19] on benchmark data sets.

Proposal to ground-truth matching in CNN-based models. Comparing the prediction to ground truth is one of the most crucial steps in training an object detector. In the IoU-based matching, all proposals with an IoU greater than a threshold with a ground truth box are considered positive in two-stage models [13,21,31]. A positive gradient is propagated for all the matched proposals, while a negative gradient is propagated for all unmatched ones. The challenges of IoU-based techniques are overcome by ATSS [42] and OTA-based [11] assignment techniques, which use adaptive anchors and global perspectives to perform assignments.

Matching strategy in transformers. DETR [4] computes a cost matrix, and uses Hungarian algorithm to compute one-to-one matching between proposals and ground truth. DN-DETR [18] and DINO [39] showed one-to-one matching causes training instability and proposed denoising techniques to stabilize training and accelerate convergence. $\mathcal{H}$-DETR [15] utilized many-to-one matching strategy to enhance performance and convergence time. They proposed repetition of ground truth, followed by Hungarian matching for each to simulate many-to-one matching. In effect, they enforce k matches for each ground truth box, even if no good matching proposals are present(refer Fig. 1b). On

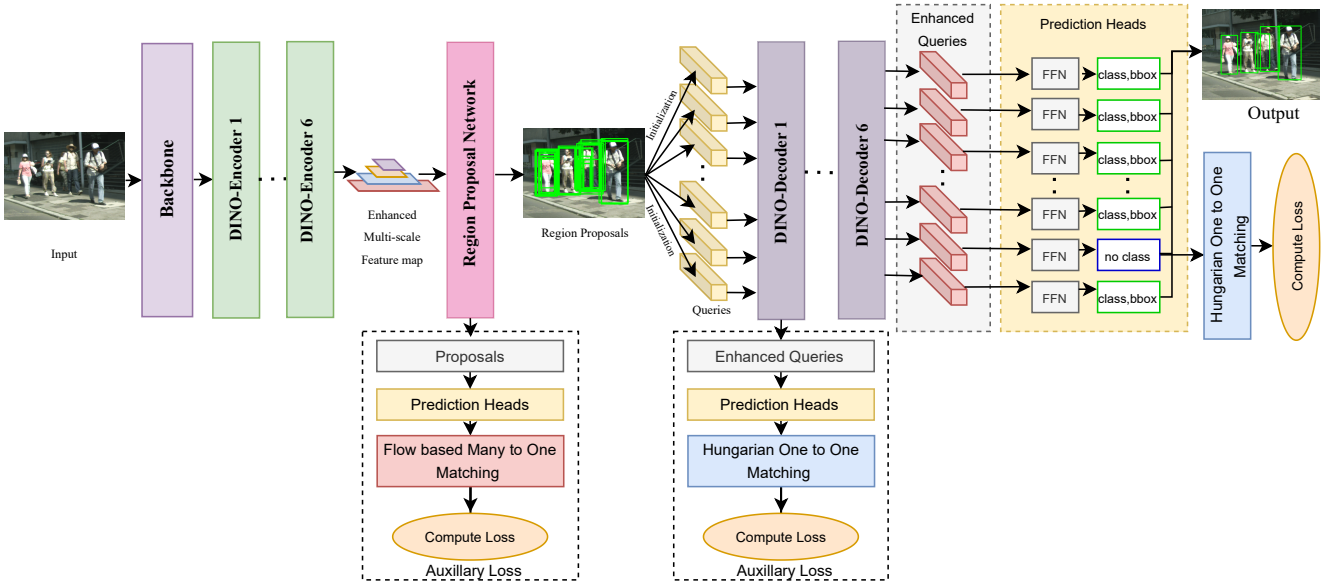

Figure 2. The overall flow of our method for a single pass during training. The Auxillary loss shown at the Region Proposal Network highlights our many-to-one matching.

the other hand, we propose a principled technique that adaptively matches between 1 to k proposals based on matching quality.

3. Proposed Methodology

Transformer based pedestrian detection. Current pedestrian models using CNN backbones follow the conventional object detection models, and require anchors for optimal performance, besides giving multiple box predictions for each ground truth. The duplicate predictions are removed using NMS before giving the final output. On the other hand, limitations identified with using anchors, and NMS step in classical object detection, have encouraged researchers to develop new transformer based models which do not have such requirements. The models typically contain an encoder and a decoder module. The encoder learns representation for the decoder, and also feeds into a module which generates proposal boxes along with their confidence. We will refer to this module as Region Proposal Network (RPN). Typically one picks top n proposals from the RPN output based on their confidence and initialize decoder queries with these. Recent works have also proposed enforcing diversity in selecting these proposals [43]. The initialized queries are passed through several layers in the decoder, where the coordinates get refined, and the self-attention among queries lead to NMS like behavior, which reduces the confidence of duplicate queries. Hence, from the final layer, one can simply pick top queries based on their confidence, and no NMS is required. Fig. 2 shows the overall flow of the baseline DINO [39] model with the proposed matching strategy at the RPN.

Gradient scaling in transformer RPN. It has been reported [14] that transformer based models give inferior performance for the pedestrian detection. In this paper, we investigate the reasons for lower performance. Our first key observation is that there are usually very small changes in box coordinates from the initialized query to the final predictions. Hence, the lower performance can be attributed to bad query initialization or bad proposals from RPN. We observe that current transformer based models train RPN by choosing one proposal per GT from the predicted set (one-to-one matching), and backpropagate positive gradient for it. All the other proposals, no matter however close to the chosen one, receive negative gradient. It is easier to understand how this might hinder learning using a simple toy example as given below [43]. Consider m similar proposals around a pedestrian in an image. Let $\{p_1, \ldots, p_m\}$ denote the confidence associated with each of the m proposals. Without loss of generality, let us assume binary cross-entropy loss while training RPN. Using the above one-to-one matching strategy, the computed loss is:

$$L_{\text{transformer}} = -\log p_1 - \sum_{i=2}^{m} \log(1 - p_i). \tag{1}$$

Recall that, in CNN models all m proposals would have been identified as positives, giving the following loss:

$$L_{\text{cnn}} = -\sum_{i=1}^{m} \log p_i. \tag{2}$$

Since all the k proposal are similar, we assume $p_1 = p_2 = \cdots = p_m = p$, and the ratio of two gradients becomes:

$$\eta = \frac{\partial L_{\text{transformer}}}{\partial p} / \frac{\partial L_{\text{cnn}}}{\partial p} = \frac{1 - mp}{m(1-p)} \tag{3}$$

Clearly, we see that the gradient is getting scaled down when $0 < p < \frac{1}{m}$ as η will be in the range of $(0, 1)$ and a negative training occurs when $\frac{1}{m} < p < 1$ as η will be in the range $(-\infty, 0)$

Naive solution strategy. Clearly, one naive solution strategy is to go the CNN way, and choose all good proposals corresponding to a ground truth for transformers as well. However, this backfires. Backpropagating positive gradients from overlapping proposals, encourages RPN to generate more such, and in-turn increases the load on decoder to reject overlapping ones, as there is no NMS step after decoder output. Further since top few proposals are used to initialize queries, getting high confidence for many overlapping boxes, may starve the difficult samples, for which one may get only a few lesser confidence boxes which will get eliminated before query initialization. Empirically also, we observe higher false positives, and more missed detections with such an approach.(refer Tab. 6)

3.1. Our proposal: many-to-one matching

We propose a middle ground between existing approaches adopted in CNN and transformer models, by allowing a GT to match with *upto* k proposals. Eq. (1) and Eq. (3) now becomes:

$$L_{\text{ours}} = -\sum_{i=1}^{k} \log p_i - \sum_{i=k+1}^{m} \log(1 - p_i), \tag{4}$$

$$\eta = \frac{\partial L_{\text{ours}}}{\partial p} / \frac{\partial L_{\text{cnn}}}{\partial p} = \frac{1 - \beta p}{\beta(1-p)}. \tag{5}$$

Here $\beta = m/k$. In the scenario, that $m = k$, there is no gradient scaling in our case, whereas, our method reduces to one-to-one matching if $k = 1$. We use $k = 6$ in our experiments, as determined by our ablation study.

Cost matrix generation. To perform a match between a proposal and a ground truth, the quality of the match between them must be quantified by a cost value. Let $\{q_1, q_2, ...q_n\}$ denote a set of n proposals and $\{p_1, p_2, ...p_n\}$ denote the confidence value corresponding to each proposal. Let $\{g_1, g_2, ...g_m\}$ denote a set of m ground-truths. We compute the matching cost between all the proposals and ground truth pedestrians, resulting in the formation of a matrix $C = [c_{i,j}]_{n \times m}$, where $c_{i,j}$ is the cost incurred if proposal q_i is assigned/matched to ground-truth g_j, and is computed as follows:

$$c_{i,j} = \lambda_{\text{class}} \mathcal{C}_{\text{class}}(q_i, g_j) + \mathcal{C}_{\text{bbox}}(q_i, g_j). \tag{6}$$

Here, $\mathcal{C}_{\text{class}}$ is the focal loss [21], computed as:

$$\mathcal{C}_{\text{class}}(q_i, g_j) = \alpha(1 - p_i)^{\gamma} \log p_i + (1 - \alpha) p_i^{\gamma} \log(1 - p_i),$$

where γ is the focusing parameter, and α is the weighting factor. Further, $\mathcal{C}_{\text{bbox}}$ is the weighted sum of L1 norm and GIoU [32] cost between the proposal and ground-truth bounding box represented by b_{q_i} and b_{g_j} respectively. The loss is computed as:

$$\mathcal{C}_{\text{bbox}}(q_i, g_j) = \lambda_{\text{L1}} \left\| b_{q_i} - b_{g_j} \right\|_1 + \lambda_{\text{GIoU}} \, \text{GIoU}(b_{q_i}, b_{g_j}),$$

where λ_{class}, λ_{L1}, and λ_{GIoU} are hyper-parameters fixed at 1, 5, and 2 respectively in our implementation.

Naive optimization strategy. Having proposed many-to-one matching and the cost matrix, one can use multiple techniques to compute the required matching. Models such as $\mathcal{H}$-DETR [15] and Group-DETR [5] use heuristics to simulate many-to-one matching using standard Hungarian matching algorithm for faster training convergence. In these works, ground truths are repeated k times, and then a traditional Hungarian matching algorithm is used to find matches. However, this constrains every ground truth to match with *exactly* k proposals, even if there are fewer relevant proposals. As shown in Fig. 1b, this can result in erroneous matches between proposals and the ground truth, diminishing the training efficiency and performance of the model Tab. 6. In contrast, we propose a principled technique to compute *upto* k matches, which can match anywhere between 1 to k proposals to the ground truth depending upon the quality of the match.

3.2. Proposed many-to-one matching algorithm

We propose a min-cost-flow based matching algorithm which assures that at least one match exists per ground truth and only attempts to match when a good match (as per the cost matrix) really exists (upto k). Below we describe the formulation and optimization details for the algorithm.

Notation. Let $G = (V, E)$ be a directed graph with vertices V and edges E. For every edge $(a, b) \in E$, $f(a, b) \in \mathbb{R}$ denotes the flow in it. Further, $l(a, b) \in \mathbb{R}$ and $u(a, b) \in \mathbb{R}$ denote the lower and upper flow capacities in edge (a, b), such that $l(a, b) \leq u(a, b)$. In an abuse of notation, we use $f(a)$ to denote net flow on a vertex $a \in V$ such that:

$$f(a) = \underbrace{\sum_{\{b:(a,b)\in E\}} f(a, b)}_{\text{outgoing flow from a}} - \underbrace{\sum_{\{b:(b,a)\in E\}} f(b, a)}_{\text{incoming flow to a}}. \tag{7}$$

For every edge, $(a, b) \in E$, we also define a cost $c(a, b) \in \mathbb{R}$ such that penalty of flow $f(a, b)$ in edge (a, b) is $f(a, b)c(a, b)$.

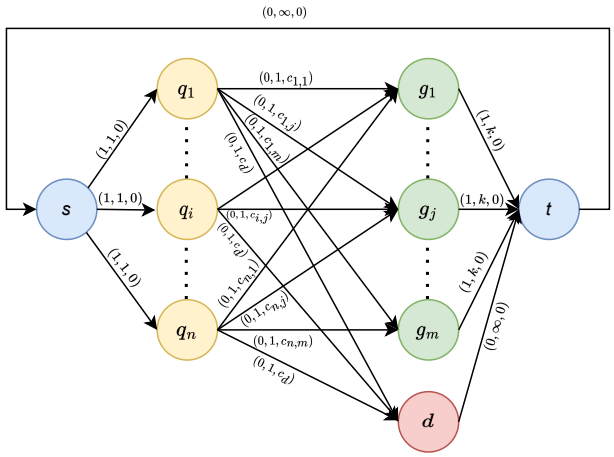

Figure 3. Constructed graph to perform our proposed many-to-one matching. The tuple (x,y,z) on every edge denotes lower-capacity x, upper-capacity y and the cost z for every edge. The node colored in red indicates the dummy ground truth node. Matching a proposal with a dummy node indicates that there is no object corresponding to it.

3.3. Minimum-cost flow objective

Given a graph G as described above, the min cost-flow problem solves the following constrained optimization:

$$\min \sum_{(a,b)\in E} c(a,b)f(a,b) \tag{8}$$

$$\text{s.t.} \quad l(a,b) \leq f(a,b) \leq u(a,b), \quad \forall (a,b) \in E \tag{9}$$

$$\text{and} \quad f(a) = 0, \quad \forall a \in V. \tag{10}$$

The first constraint ensures that the flow satisfies minimum and upper bounds at every edge. The second constraint ensures that flow is conserved at every vertex, that is, the difference between outgoing and incoming flow is zero at every vertex.

3.3.1 Many-to-one matching as minimum cost flow

To map many-to-one matching to a minimum cost flow problem, we create graph $G = (V, E)$ as follows.

1. The vertex set $V = \{q_1, \ldots, q_n, g_1, \ldots, g_m, s, t, d\}$. Here q_i denotes the vertex corresponding to the i^{th} proposal, g_j is the vertex corresponding to the j^{th} ground-truth, and s, t, and d denote the source, sink and the dummy ground truth. The purpose of s, t, and d will be described below.

2. The edge set E contains the following directed edges.
   (a) We add a directed edge (q_i, g_j) between each proposal and ground truth node with cost $c_{i,j}$ as given by Eq. (6). We define lower capacity $l(q_i, g_j) = 0$, and $u(q_i, g_j) = 1$ for each such edge.
   (b) We add a directed edge (q_i, d) between each proposal node q_i and dummy ground truth vertex d. The cost of each such edge $c(q_i, d) = c_d$. The value of c_d has been empirically estimated as discussed in Tab. 7. The minimum and maximum flow in the edges are set as: $l(q_i, d) = 0$, and $u(q_i, d) = 1$, $\forall i = \{1, \ldots, n\}$.
   (c) We add directed edges (s, q_i) from source to each proposal node with $c(s, q_i) = 0$, $l(s, q_i) = 1$ and $u(s, q_i) = 1$, $\forall i = \{1, 2...n\}$. This ensures that each proposal is matched to atleast and at-most one of the real or dummy ground truth node.
   (d) We add directed edges (g_j, t) from each ground truth node to sink with $c(g_j, t) = 0$, $l(g_j, t) = 1$, and $u(g_j, t) = k$, $\forall j = \{1, 2...m\}$. This ensures that each real ground truth node is matched to atleast 1 and at-most k proposal nodes.
   (e) We add a directed edge (d, t) between the dummy ground truth node and sink node such that $c(d, t) = 0$, $l(d, t) = 0$ and $u(d, t) = \infty$, making it practically an unconstrained edge.
   (f) Finally we add another directed, and unconstrained edge (t, s) between sink and source node with $c(d, t) = 0$, $l(d, t) = 0$ and $u(d, t) = \infty$.

Fig. 3 shows the constructed graph. To estimate the minimum cost flow in G, we employ the push-relabel algorithm [12]. After obtaining the flow $f(q_i, g_j)$ $\forall i \in \{1, 2...n\}$ and $\forall j \in \{1, 2, .., m\}$, matching between the proposals and ground truth boxes can be accomplished by selecting those edges (q_i, g_j) where $f(q_i, g_j) = 1$. All those proposals that match with the dummy ground truth will have a $f(q_i, d) = 1$, and these matches are ignored. The cost c_d acts as the threshold to assess the quality of match. All edges with costs exceeding c_d will be matched to a dummy ground truth instead of the real ground truth, thus preventing poor matches. Since the lower capacity of the edges $(g_j, t) = 1$ $\forall j = \{1, 2...m\}$, it ensures that there is at least one match for each ground-truth.

3.4. Complexity Analysis

Our approach introduces no added complexity during inference. When training, the push-relabel algorithm [12] has a time complexity of $O(|V|^3 \log(|V|C))$, with C being the maximum cost in the graph and $|V|$ the number of vertices. Typically, the $\log(|V|C)$ value falls within $[6, 12]$. This results in a time complexity similar to the Hungarian algorithm,$O(|V|^3)$. Our empirical findings confirm this, showing a 5.8% increase in training time over the baseline on the ECP dataset.

In supplementary, we give the construction and proof that the matching produced by our method is equivalent to

that by the Hungarian algorithm for the special case of one-to-one matching.

3.5. Loss function

We observe that proposal predictions from baseline DINO model are misaligned with respect to the ground truth. That is predictions with "high IoU" with the ground truth have a low confidence score, whereas predictions with a moderate IoU have a high confidence score. We hypothesize that the independent classification cost ($\mathcal{C}_{\text{class}}$) and the bounding box cost ($\mathcal{C}_{\text{bbox}}$) used in the Eq. (6) during matching is responsible for the issue. The matching strategy used in DETR-like models may produce a prediction and ground truth match with the lowest matching cost, which never guarantees that the prediction has the highest confidence score and the best IoU value. Hence, we use an IoU-aware classification loss [3], which also includes an IoU value between the matched prediction and the ground truth during the computation of the classification loss. It penalizes predictions that have a match with low IoU and high classification score. Let $Q = \{q_1, \ldots, q_n\}$ denote a set of predictions with confidence scores $\{p_1, \ldots, p_n\}$ and $S \subseteq Q$ be the set of predictions that are matched with the ground truths. The proposed loss is defined as follows:

$$\mathcal{L} = \sum_{q_i \in S} \text{BCE}(p_i, t_i) + \sum_{q_i \in Q \setminus S} p_i^2 \, \text{BCE}(p_i, 0). \quad (11)$$

Here BCE is the binary cross-entropy loss and t_i is computed as follows:

$$t_i = p_i^w \, u_i^{1-w}. \quad (12)$$

Here u_i is the IoU score between the prediction q_i and its matched ground truth and w is a hyperparameter which lies in the range $[0, 1]$. In effect, t_i is the weighted geometric mean of p_i and u_i. The value of w is empirically selected as 0.25. More discussion on the topic is done in the supplementary material.

4. Dataset and Evaluation Methodology

Datasets used. We employed diverse datasets containing daytime autonomous driving images, as shown in Tab. 1. The Euro City Persons (ECP) dataset offers images from 12 European countries, both during the day and night. For consistency with other baselines, we focused on daytime scenes. The City-Persons (CP) dataset features daytime images from 27 German cities, with slightly fewer pedestrians and scene variations compared to ECP. The Caltech Pedestrian dataset, with $42,872$ daylight images, has lower resolution and less densely distributed pedestrians than ECP and CP. The TJU-Ped-Traffic and TJU-Ped-Campus datasets portray pedestrians during daytime across various seasons, weather, and lighting. We use the validation set for comparison on ECP, TJU-Ped, and Citypersons datasets. For Caltech, we use the provided test set.

Dataset	Images	APPI	Time	Resolution
Caltech Ped. [9]	42,782	0.32	day	640×480
Citypersons [41]	2,975	6.47	day	2048×1024
ECP [1]	21,795	9.2	day,night	1920×1024
TJU-Ped-Traffic [28]	13,858	2.0	day	1624×1200
TJU-Ped-Campus [28]	39,727	5.9	day	1624×1200

Table 1. Summary of Benchmark Pedestrian Datasets. APPI is acronym for average pedestrians per image.

Setting	Height	Visibility
Reasonable	$[50, \infty]$	$[0.65, \infty]$
Small	$[50, 75]$	$[0.65, \infty]$
Heavy	$[50, \infty]$	$[0.2, 0.65]$
All	$[20, \infty]$	$[0.2, \infty]$

Table 2. Summary of Evaluation Setting [13]. Height is the height of the bounding box in pixels, Visibility is the ratio of the visible region of an object within the bounding box to its total area.

Training details. We have implemented the experimental design utilized by DINO [39]. We use pre-trained weights of the DINO model with the Swin-L [27] backbone, trained on MS-COCO dataset [22]. While training, we used the loss function described in Sec. 3.5, as the classification loss and the L1 norm and GIoU Loss as the regression loss. The network is optimized using AdamW with a learning rate of 1×10^{-4} and a weight decay of 1×10^{-4}. We have used four A100 GPUs to train the model for 36 epochs with a batch size of 8. The maximum number of matches per ground truth (k) has been set to 6 with a dummy ground truth cost of $c_d = 4.5$.

Evaluation metric. All values are reported using the evaluation metric $Log\ Average\ Miss\ Rate$ (written as MR^{-2} or miss rate). MR^{-2} is determined by calculating the log average of miss rates at nine distinct thresholds uniformly distributed in the log space $[10^{-2}, 10^{0}]$. All the numbers are reported as $\%\ MR^{-2}$. We report the MR^{-2} values for the reasonably-sized, small-sized, and heavily occluded objects. The lower MR^{-2} is, the better. Tab. 2 shows the different evaluation settings. The height attribute specifies the height of the bounding box in pixels, while the Visibility attribute specifies the ratio of the areas of the visible region of the object to the total area of the bounding box.

5. Results and Discussions

Comparison with the SOTA. Tab. 3 demonstrates that our model outperforms other SOTA across five different pedes-

Method	Reasonable	Small	Heavy
Euro City Persons [1]			
YOLOv3 [30]	8.5	17.8	37.0
FRCNN [31]	7.3	16.6	52.0
Pedestron [13]	6.6	13.6	33.3
F2DNet [16]	6.1	10.7	28.2
LSFM [17]	4.7	**9.9**	23.8
Ours	**3.7**	10.4	**19.9**
TJU-Pedestrian-Traffic [28]			
F2DNet [16]	21.6	26.3	62.6
CrowdDet [6]	20.8	–	61.2
EGCL [23]	19.7	–	60.1
Pedestron [13]	18.9	**24.0**	56.3
LSFM [17]	18.7	24.9	56.2
Ours	**17.4**	24.7	**52.68**
TJU-Pedestrian-Campus [28]			
RetinaNet [21]	34.73	82.99	71.31
DeFCN [37]	32.1	62.7	72.7
FCOS [35]	31.89	69.04	81.28
OPL [33]	31.5	61.7	72.4
FPN [20]	27.92	67.52	73.14
CrowdDet [6]	25.73	–	66.38
EGCL [23]	24.84	–	65.27
Ours	**21.83**	**37.04**	**57.08**
Citypersons [41]			
Pedestron [13]	11.2	14	37
CSP [13, 26]	11	16	49.3
PRNet [34]	10.8	–	42
APD [40]	8.8	–	46.6
F2DNet [16]	8.7	11.3	32.6
LSFM [17]	8.5	**8.8**	31.9
Ours	**8.3**	15.56	**27.07**
Caltech [9]			
Pedestron [13]	6.2	7.4	55.3
ALFNet [25]	6.1	7.9	51
AR-Ped [2]	4.4	–	48.8
F2DNet [16]	2.2	**2.5**	38.7
LSFM [17]	3.1	3.4	**35.8**
Ours	**2.0**	2.8	38.6

Table 3. Our results on various benchmark datasets based on MR^{-2} (lower is better). **Bold** indicates the best, and underline indicates the second best methods.

trian benchmark datasets. One can observe an average enhancement of 1.3% and 4.1% in the MR^{-2} for reasonable and heavy occlusion, respectively.

Cross Dataset evaluation. We conduct a cross-dataset evaluation to demonstrate our model's generalizability to images from a different dataset. Tab. 4 demonstrates our superiority in cross-dataset evaluation compared to SOTA.

Method	Train	Test	Reasonable	Small	Heavy
CSP [13, 26]	ECP	CP	11.5	16.6	38.2
Pedestron [13]	ECP	CP	10.9	11.4	40.9
F2DNet [16]	ECP	CP	10.1	12.1	**36.4**
LSFM [17]	ECP	CP	9.4	**11.1**	37.8
Ours	ECP	CP	**9.2**	12.4	37.2
F2DNet [16]	ECP	Caltech	16.9	21.5	41.3
LSFM [17]	ECP	Caltech	13.1	16.3	33.1
CSP [13, 26]	ECP	Caltech	10.4	13.7	31.3
Pedestron [13]	ECP	Caltech	8.1	9.6	29.9
Ours	ECP	Caltech	**8.1**	**7.25**	**21.8**
LSFM [17]	CP	Caltech	11.7	15.6	37.4
F2Dnet [16]	CP	Caltech	11.3	13.7	32.6
CSP [13, 26]	CP	Caltech	10.1	13.3	34.4
Pedestron [13]	CP	Caltech	8.8	9.8	28.8
Ours	CP	Caltech	**8.52**	**8.5**	**23.41**
CSP [13, 26]	CP	ECP	19.6	51	56.4
Pedestron [13]	CP	ECP	17.2	40.5	49.3
LSFM [17]	CP	ECP	17	42.1	49.6
F2DNet [16]	CP	ECP	**11.6**	**14.7**	40
Ours	CP	ECP	17.4	57.2	**38.33**

Table 4. Our results on cross dataset validation based on MR^{-2}(lower is better). **Bold** indicates the best, and underline indicates the second best methods. In this analysis, the model is trained on one dataset and evaluated on another dataset's images.

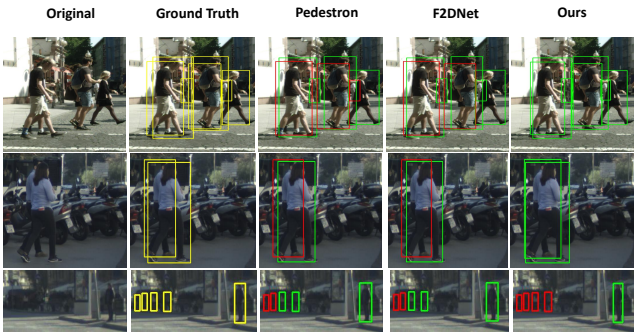

Figure 4. Visual comparison of the outputs from our model with the existing SOTAs.

Qualitative Analysis. Fig. 4 shows the visual comparison of our model with the other SOTA models. Yellow boxes represent ground truth, red boxes signify missed predictions, and green boxes show actual model predictions. It can be observed from the images that our model is highly accurate in detecting pedestrians in challenging high-occlusion scenarios. However, it can be observed that our model performs slightly below par when detecting small pedestrians.

Ablation studies. We perform various ablation studies to understand the contributions made by each component in our model. Unless explicitly mentioned all the ablations are performed on the Euro City Persons Dataset [1].

Components		Categories		
Proposed Loss	Flow-based Many to one	Reasonable	Small	Heavy
no	no	4.25	11.25	20.44
yes	no	4.1	11	20.5
yes	yes	**3.7**	**10.4**	**19.9**

Table 5. Ablation study to understand the impact of proposed matching. We observe that incorporating our flow-based matching resulted in a better performance highlighting the significance of good matching between the proposal and the ground truth.

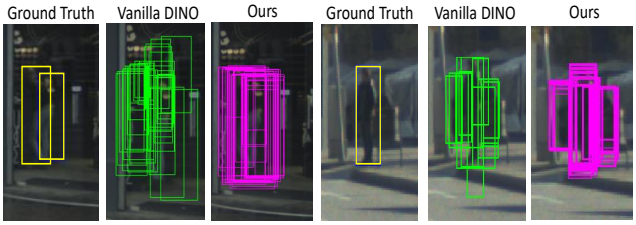

Figure 5. The qualitative analysis of the region proposals proposed by our method in comparison to that of the vanilla DINO.

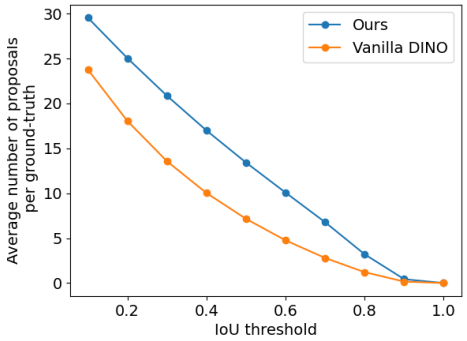

Figure 6. A plot showing the average increase in the number of proposals per ground truth compared with that of vanilla DINO.

- Tab. 5 illustrates the significance of using proposed loss function (Sec. 3.5), and many-to-one matching.
- Fig. 5 illustrates the comparison between the region proposals generated per object at the region proposal layer. It can be observed that the introduction of many-to-one matching contributes to the generation of improved and more precise object-centered proposals. Fig. 6 demonstrates that our method consistently has more proposals per ground truth on average. At an IoU of 0.5, there is an increase of 100% of region proposals per ground truth.
- Tab. 6 shows impact of using different matching techniques on the detection performance.
- Tab. 7 shows results of experiments conducted at various threshold values used for matching with dummy ground truth node. We obtain the best results for 4.5.

Matching Algorithm	Reasonable	Small	Heavy
Fixed IoU based [31]	9.05	17.09	28.55
ATSS [42]	6.10	14.22	26.25
OTA [11]	4.73	12.36	21.55
Hungarian with GT rep. [15]	4.1	11.2	21.01
Flow-based (proposed)	**3.7**	**10.4**	**19.9**

Table 6. Ablation study to understand the impact of different techniques to match proposal to ground truth at RPN. The proposed flow-based matching outperforms the Hungarian with repeated ground-truth matching, the Fixed IoU-based, the ATSS, and the OTA strategies.

Max matches Per Ground truth	Dummy node cost	Reasonable	Small	Heavy
6	2.5	4.18	10.9	21.01
6	3.5	3.9	10.8	20.6
6	4.5	**3.7**	**10.4**	**19.9**
6	5.5	4.2	10.8	21.6
6	6.5	4.1	11.4	20.8

Table 7. Ablation study varying the dummy weight.

Max matches Per Ground truth	Dummy node cost	Reasonable	Small	Heavy
4	4.5	4.1	10.7	20.7
5	4.5	4.0	10.6	20.4
6	4.5	**3.7**	**10.4**	**19.9**
7	4.5	3.9	10.7	20.0

Table 8. Ablation study to understand the impact of variation of the maximum number of matches per ground-truth.

- Tab. 8 shows the result of changing k, that is number of maximum matches per ground truth.
- We showcase the method's generalizability on the MS COCO [22] dataset. Due to space constraints, the results are summarized in the supplementary material

6. Conclusion

In this paper, we identified the problem of one-to-one matching between proposals and ground truth as the cause of the lower performance of transformer-based pedestrian detection techniques. While ad-hoc solutions such as duplicating ground truth k times exist, they enforce matching exactly k proposals to ground truth, leading to the misclassification of incorrect proposals as positive proposals. Instead, we gave a principled solution using minimum cost flow-based many-to-one matching that solves this problem. Contrary to the recent works which claim that CNNs perform better than transformers on pedestrian detection, using our approach, we establish a new transformer-based state-of-the-art on multiple benchmark datasets.

Acknowledgement. This work has been partly supported by the funding received from DST through the ICPS scheme and DRDO-YSL-AI Lab.

References

[1] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019. 6, 7

[2] Garrick Brazil and Xiaoming Liu. Pedestrian detection with autoregressive network phases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7231–7240, 2019. 7

[3] Zhi Cai, Songtao Liu, Guodong Wang, Zheng Ge, Xiangyu Zhang, and Di Huang. Align-detr: Improving detr with simple iou-aware bce loss. *arXiv preprint arXiv:2304.07527*, 2023. 6

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 213–229. Springer, 2020. 1, 2

[5] Qiang Chen, Xiaokang Chen, Gang Zeng, and Jingdong Wang. Group detr: Fast training convergence with decoupled one-to-many label assignment. *arXiv preprint arXiv:2207.13085*, 2022. 4

[6] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12214–12223, 2020. 2, 7

[7] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. Ieee, 2005. 2

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[9] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: A benchmark. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 304–311, 2009. 6, 7

[10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009. 2

[11] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 303–312, 2021. 2, 8

[12] Andrew V Goldberg and Robert E Tarjan. A new approach to the maximum-flow problem. *Journal of the ACM (JACM)*, 35(4):921–940, 1988. 5

[13] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11328–11337, 2021. 1, 2, 6, 7

[14] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Pedestrian detection: Domain generalization, cnns, transformers and beyond. *arXiv preprint arXiv:2201.03176*, 2022. 2, 3

[15] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19702–19712, 2023. 1, 2, 4, 8

[16] Abdul Hannan Khan, Mohsin Munir, Ludger van Elst, and Andreas Dengel. F2dnet: Fast focal detection network for pedestrian detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4658–4664. IEEE, 2022. 1, 2, 7

[17] Abdul Hannan Khan, Mohammed Shariq Nawaz, and Andreas Dengel. Localized semantic feature mixers for efficient pedestrian detection in autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5476–5485, 2023. 1, 2, 7

[18] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 2

[19] Matthieu Lin, Chuming Li, Xingyuan Bu, Ming Sun, Chen Lin, Junjie Yan, Wanli Ouyang, and Zhidong Deng. Detr for crowd pedestrian detection. *arXiv preprint arXiv:2012.06785*, 2020. 2

[20] 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*, pages 2117–2125, 2017. 7

[21] 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*, pages 2980–2988, 2017. 2, 4, 7

[22] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6, 8

[23] Zebin Lin, Wenjie Pei, Fanglin Chen, David Zhang, and Guangming Lu. Pedestrian detection by exemplar-guided contrastive learning. *IEEE transactions on image processing*, 2022. 2, 7

[24] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022. 1

[25] 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. 1, 2, 7

[26] 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/CVF conference on computer vision and pattern recognition*, pages 5187–5196, 2019. 1, 2, 7

[27] 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. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6

[28] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanqing Sun, and Jinfeng Gong. Tju-dhd: A diverse high-resolution dataset for object detection. *IEEE Transactions on Image Processing*, 30:207–219, 2020. 6, 7

[29] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 1

[30] 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*, pages 779–788, 2016. 2, 7

[31] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 2, 7, 8

[32] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019. 4

[33] Xiaolin Song, Binghui Chen, Pengyu Li, Jun-Yan He, Biao Wang, Yifeng Geng, Xuansong Xie, and Honggang Zhang. Optimal proposal learning for deployable end-to-end pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3250–3260, 2023. 7

[34] Xiaolin Song, Kaili Zhao, Wen-Sheng Chu, Honggang Zhang, and Jun Guo. Progressive refinement network for occluded pedestrian detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 32–48. Springer, 2020. 2, 7

[35] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 7

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[37] Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with fully convolutional network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15849–15858, 2021. 7

[38] Wenhao Wang. Adapted center and scale prediction: more stable and more accurate. *arXiv preprint arXiv:2002.09053*, 2020. 2

[39] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Harry Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *International Conference on Learning Representations*, 2022. 1, 2, 3, 6

[40] Jialiang Zhang, Lixiang Lin, Jianke Zhu, Yang Li, Yun-chen Chen, Yao Hu, and Steven CH Hoi. Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*, 23:3085–3097, 2020. 7

[41] 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. 6, 7

[42] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9759–9768, 2020. 2, 8

[43] Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7329–7338, 2023. 3

[44] 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. 1