What Makes for End-to-End Object Detection?

Peize Sun [1] **Yi Jiang** [2] **Enze Xie** [1] **Wenqi Shao** [3] **Zehuan Yuan** [2] **Changhu Wang** [2] **Ping Luo** [1]

Abstract

Object detection has recently achieved a breakthrough for removing the last one non-differentiable component in the pipeline, Non-Maximum Suppression (NMS), and building up an end-to-end system. However, what makes for its one-to-one prediction has not been well understood. In this paper, we first point out that *one-to-one positive sample assignment* is the key factor, while, one-to-many assignment in previous detectors causes redundant predictions in inference. Second, we surprisingly find that even training with one-to-one assignment, previous detectors still produce redundant predictions. We identify that *classification cost* in matching cost is the main ingredient: (1) previous detectors only consider location cost, (2) by additionally introducing classification cost, previous detectors immediately produce one-to-one prediction during inference. We introduce the concept of *score gap* to explore the effect of matching cost. Classification cost enlarges the score gap by choosing positive samples as those of highest score in the training iteration and reducing noisy positive samples brought by only location cost. Finally, we demonstrate the advantages of end-to-end object detection on crowded scenes.

1. Introduction

Object detection is one of the fundamental tasks in the computer vision area and enables numerous downstream applications. It aims at localizing a set of objects and recognizing their categories in an image. The development of object detection pipeline (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015; Cai & Vasconcelos, 2018; Redmon et al., 2016; Liu et al., 2016; Lin et al., 2017b; Tian et al., 2019; Zhou et al., 2019; Carion et al., 2020) is a route to remove manually-designed components and towards end-to-end system.

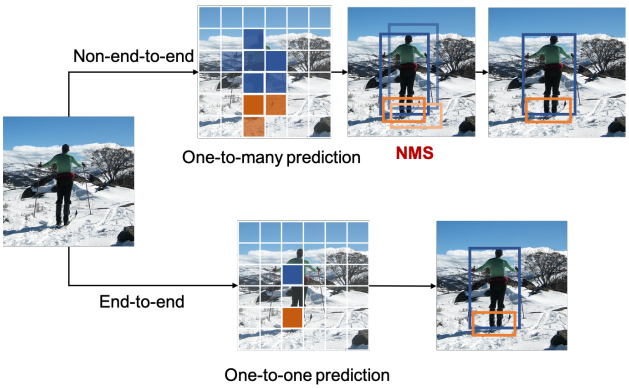


Figure 1. **End-to-end object detection.** Non-end-to-end object detectors require NMS to remove redundant predictions. As the last one manually-designed component in the object detection pipeline, non-differentiable NMS blocks setting up an end-to-end object detection system.

For decades, the sample in object detection is box candidates. In classical computer vision, the classifier is applied on sliding windows enumerated on the image grid (Dalal & Triggs, 2005; Felzenszwalb et al., 2010; Viola & Jones, 2001). Modern detectors pre-define thousands of anchor boxes on the image grid and perform classification and regression on these candidates (Girshick et al., 2014; Ren et al., 2015; Lin et al., 2017b; Redmon & Farhadi, 2017).

Despite box candidate methods dominate object detection for years, the detection performance is largely sensitive to sizes, aspect ratios, and the number of anchor boxes. To eliminate the hand-crafted design and complex computation of box candidates, anchor-free detectors (Tian et al., 2019; Zhou et al., 2019) are rising. These methods directly treat grid points in the feature map as object candidates and predict the offset from the grid point to the object box's boundaries and largely simplify the detection pipeline.

However, both box candidates and point candidates suffer from one common problem, that is, redundant and near-duplicate predictions for each object are produced, thus making non-maximum suppression(NMS) necessary post-processing in inference. Towards building up an end-to-

[1]Department of Computer Science, The University of Hong Kong [2]AI Lab, ByteDance [3]Department of Electronic Engineering, The Chinese University of Hong Kong. Correspondence to: Peize Sun <peizesun@connect.hku.hk>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

end object detection system, NMS is the last one manually-designed component in the pipeline.

Recently, attention-based detectors (Hu et al., 2018; Carion et al., 2020; Zhu et al., 2020; Sun et al., 2020a) achieve to directly output predictions without NMS. Thus far, all manually-designed components in the pipeline are removed and an end-to-end object detection system is finally set up. However, both attention-based architecture and one-to-one positive sample assignment of these detectors are brand new compared with previous methods based on box and point candidates. *It motivates us to explore what exactly makes for end-to-end object detection.*

In order to understand what enables non-redundant prediction in object detection, we study on three non-end-to-end detectors, RetinaNet (Lin et al., 2017b), CenterNet (Zhou et al., 2019), FCOS (Tian et al., 2019) and three end-to-end detectors, DETR (Carion et al., 2020), Deformable DETR (Zhu et al., 2020), Sparse R-CNN (Sun et al., 2020a). Our empirical findings show that:

- Non-end-to-end detectors assigning positive samples by one-to-many groundtruth-to-samples causes redundant predictions in inference, while, end-to-end detectors are one-to-one assigning. However, even training with one-to-one assignment, non-end-to-end detectors still produce redundant predictions.
- The lack of classification cost is the main obstacle to achieve one-to-one prediction: (1) non-end-to-end detectors only consider location cost. (2) by additionally considering classification cost, these detectors immediately produce one-to-one prediction during inference, which successfully removes NMS and achieves end-to-end detection.

Since redundant predictions are those of high classification scores, we introduce the concept of *score gap* to describe the gap between the first-highest score and the second-highest score. A sufficient requirement for end-to-end detection is that the score gap should be large enough. Assigning positive samples by only location cost cannot enlarge the score gap since it chooses positive samples as those of medium classification score in training iterations, while additionally considering classification cost leads to an enough large score gap by choosing those samples of the highest score. Moreover, we identify that positive samples chosen by only location cost introduce background-like positive samples, thus decrease the discriminative ability of the network, while classification cost could reduce these noisy samples.

We analyze the convergence properties of one-to-one positive sample assignment with classification cost using perceptron's update rule in the linearly separable setting.

End-to-end object detectors avoid NMS dilemma (Zhang et al., 2019) in crowded scenes. In CrowdHuman dataset (Shao et al., 2018), we demonstrate that end-to-end versions of RetinaNet and FCOS outperform their baseline settings by a large margin.

2. Preliminary on Object Detection

Object detection is a multi-task of localizing a set of objects and recognizing their categories in an image. For an input image of $H \times W \times 3$, the predictions are N boxes with categories of $N \times K$ and locations of $N \times 4$, where K is the number of categories and 4 is coordinates of four sides.

2.1. Pipeline

Object Candidate. Object detectors assume a region in feature map (Girshick et al., 2014; Ren et al., 2015; Cai & Vasconcelos, 2018) or a point in feature map (Redmon et al., 2016; Lin et al., 2017b; Tian et al., 2019; Zhou et al., 2019) as the object candidate. The number of object candidates is always much more than possible objects to guarantee detection recall.

Classification and Location. The classification sub-net predicts the probability of object candidate for K object categories. The location sub-net predicts the offset from each object candidate to 4 boundaries of the object box.

2.2. Training

Loss of object detection. The training loss of the object detection includes classification loss and regression loss, where regression loss is only executed on positive sample:

$$\begin{aligned} L &= \sum_{i \in \mathcal{P} \cup \mathcal{S} \setminus \mathcal{P}} L_{cls}(i) + \sum_{i \in \mathcal{P}} L_{loc}(i) \\ &= \sum_{i \in \mathcal{P}} [L_{cls}(i) + L_{loc}(i)] + \sum_{i \in \mathcal{S} \setminus \mathcal{P}} L_{cls}(i) \end{aligned} \tag{1}$$

where $\mathcal{S}$ is the set of samples, $\mathcal{P}$ is the set of positive sample, $\mathcal{S} \setminus \mathcal{P}$ is the set of negative sample, L_{cls} is classification loss between predicted category and ground-truth category, such as cross entropy loss and Focal Loss (Lin et al., 2017b), L_{loc} is location loss between sample box and ground-truth box, such as L1 loss and GIoU loss (Rezatofighi et al., 2019).

Though the training loss of object detection is well-defined, positive samples are controversial. In object detection, the annotation is bounding box and category of the object in the image, instead of object candidates. Selecting positive samples in object detection task is more complicated than image-level classification task, since positive samples and negative ones for image classification are indisputable

when the image annotation is given.

Matching cost. To better select positive samples and negative ones for object detection, matching cost is introduced to measure the distance between the sample and object. For sample i and object j, the matching cost $C_{i,j}$ is:

$$C_{i,j} = C_{cls}(i,j) + C_{loc}(i,j) \tag{2}$$

where $C_{cls}(i,j)$ is classification loss between predicted category of sample i and ground-truth category of object j, $C_{loc}(i,j)$ is location loss between sample i and ground-truth box of object j. For convenience, we call $C_{loc}(i,j)$ as location cost, and $C_{cls}(i,j)$ as classification cost.

The matching cost is not required to be strictly equal to the loss function, as long as its design is suitable to select positive samples. In fact, the matching cost only contains $C_{loc}(i,j)$ for decades before recently (Carion et al., 2020; Zhu et al., 2020; Sun et al., 2020a).

Positive sample assignment. Once the matching cost between all samples and objects j is computed, those samples below the cost threshold $\theta(j)$ will be chosen as positive samples:

$$\mathcal{P} = \{\, i \mid C_{i,j} < \theta(j), i \in \mathcal{S}\} \tag{3}$$

Many heuristic rules (Girshick et al., 2014; Cai & Vasconcelos, 2018; Tian et al., 2019; Zhou et al., 2019; Zhang et al., 2020b;a) are proposed to determine θ, which lead to one-to-many and one-to-one assignment of groundtruth-to-positive samples.

2.3. Inference.

Since the object candidates are always much more than objects in the image, the output is filtered by the score threshold to guarantee detection precision. If there remain redundant boxes, non-maximum suppression(NMS) is used to remove these redundant predictions. NMS is a heuristic manually-designed component. The box with the maximum score is selected and others neighboring boxes are eliminated.

However, non-differentiable NMS blocks the establishment of an end-to-end system. Worsely, detectors suffer from NMS dilemma in crowded scene (Zhang et al., 2019). To this end, end-to-end object detection is proposed.

3. End-to-End Object Detection

End-to-end object detection means that object detection pipeline is *without any non-differentiable component, e.g., NMS*. The input of network is the image and the output is direct predictions of classification on object categories or background and the box regression. The whole network is trained in an end-to-end manner with back-propagation.

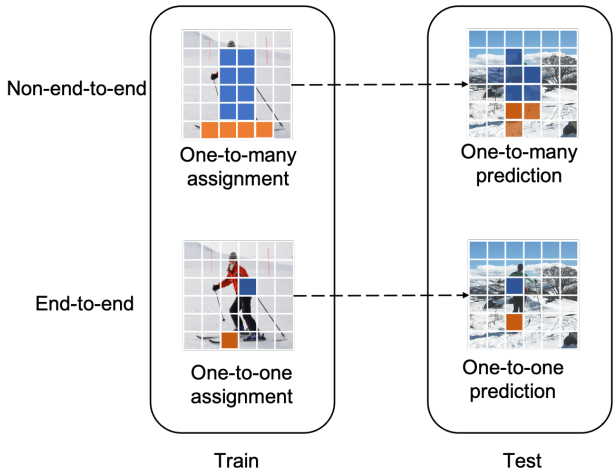


Figure 2. **Positive sample assignment.** Non-end-to-end detectors apply one-to-many positive sample assignment in training and produce one-to-many predictions in inference. While, end-to-end object detectors are one-to-one positive sample assignment and one-to-one prediction. This motivates us to apply one-to-one assignment in non-end-to-end detectors.

Detector	o2o	AP	AP(+NMS)
DETR	✓	40.0	39.9 (-0.1)
Deformable DETR	✓	44.0	43.9 (-0.1)
Sparse R-CNN	✓	45.0	44.9 (-0.1)
RetinaNet		7.7	37.4 (+29.7)
	✓	33.6	36.8 (+3.2)
CenterNet		24.9	35.0 (+10.1)
	✓	23.4	32.0 (+8.6)
FCOS		17.3	38.7 (+21.4)
	✓	34.9	37.7 (+2.8)

Table 1. **Effect of one-to-one positive sample assignment.** The detectors' original settings are highlighted by gray. "o2o" means one-to-one positive sample assignment. The top section is end-to-end detectors, which apply one-to-one assignment and don't depend on NMS. The bottom section is non-end-to-end detectors, whose original settings use one-to-many assignment and heavily rely on NMS. Training with one-to-one assignment only reduces non-end-to-end detectors' dependence on NMS to some extend, they still need NMS to further remove redundant predictions.

3.1. Experiment Setting

Detectors. We select three non-end-to-end detectors, RetinaNet (Lin et al., 2017b), CenterNet (Zhou et al., 2019), FCOS (Tian et al., 2019) and three end-to-end detectors, DETR (Carion et al., 2020), Deformable DETR (Zhu et al., 2020), Sparse R-CNN (Sun et al., 2020a).

Dataset. Our experiments are conducted on the challenging COCO benchmark (Lin et al., 2014). We use the standard COCO metrics AP of averaging over IoU thresholds. All models are trained on `train2017` split (∼118k images) and evaluated with `val2017` (5k images).

3.2. Positive Sample Assignment

One-to-many assignment. The remarkable property of non-end-to-end detectors is one-to-many positive sample assignment, as shown in Figure 2. In the training step, for one ground-truth box, any sample whose matching cost is below the cost threshold is assigned as the positive sample. It always causes multiple samples in the feature maps to be selected as positive samples. As a result, in the inference step, these detectors produce redundant predictions.

One-to-one assignment. On the contrary, end-to-end detectors apply one-to-one assignment during the training step. For one ground-truth box, only one sample with the minimum matching cost is assigned as the positive sample, others are all negative samples. The positive sample is usually selected by bipartite matching (Kuhn, 1955) to avoid sample conflict, *i.e.*, two ground-truth boxes share the same positive sample.

As shown in Table 1, end-to-end detectors, including DETR, Deformable DETR and Sparse R-CNN, apply one-to-one assignment and eliminate NMS. Therefore, an intuitive idea to transform non-end-to-end detectors become end-to-end is to replace one-to-many assignment with one-to-one assignment. Specifically, RetinaNet chooses the positive sample as the anchor that has largest IoU with the ground-truth box, CenterNet chooses the grid point in the feature map that has the nearest distance to the ground-truth box center, while FCOS chooses from the pre-defined layer in feature pyramids (Tian et al., 2019).

However, one-to-one assignment only reduces the dependence on NMS to some extend, non-end-to-end detectors still need NMS to further remove redundant predictions. For example, NMS could further improve one-to-one assignment version of RetinaNet, CenterNet and FCOS by 3.2 AP, 8.6 AP and 2.8 AP, respectively, as shown in Table 1.

Conclusion 3.1 *Even replacing one-to-many assignment to one-to-one assignment in training, non-end-to-end detectors still produce redundant predictions in inference.*

Experiments on positive sample assignment demonstrate that one-to-one assignment is necessary but not sufficient for end-to-end object detection. We further delve into the compositions of matching cost.

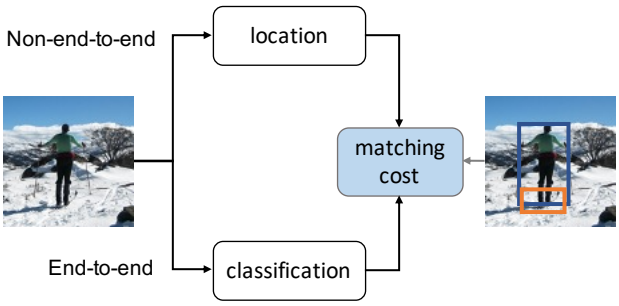


Figure 3. **Matching cost.** Non-end-to-end object detectors assign positive samples by only location cost, while end-to-end detectors additionally consider classification cost.

Detector	loc. pre-def.	loc. pred.	cls.	AP	AP(+NMS)
DETR		✓		12.6	23.6 (+11.0)
		✓	✓	40.0	39.9 (-0.1)
Deformable DETR		✓		12.0	23.8 (+11.8)
		✓	✓	44.0	43.9 (-0.1)
Sparse R-CNN		✓		20.1	33.1 (+13.0)
		✓	✓	45.0	44.9 (-0.1)
RetinaNet + o2o	✓			33.6	36.8 (+3.2)
	✓		✓	36.0	36.2 (+0.2)
		✓	✓	37.5	37.5 (+0.0)
CenterNet + o2o	✓			23.4	32.0 (+8.6)
	✓		✓	33.3	33.2 (-0.1)
		✓	✓	34.9	34.8 (-0.1)
FCOS + o2o	✓			34.9	37.7 (+2.8)
	✓		✓	35.9	36.1 (+0.2)
		✓	✓	38.9	38.9 (+0.0)

Table 2. **Effect of classification cost.** The detectors' original settings are highlighted by gray. "o2o" means one-to-one positive sample assignment. "loc." means location cost. "cls." means classification cost. "pre-def." and "pred." are pre-defined location cost and predicted location cost, illustrated in 3.3. All detectors apply one-to-one positive sample assignment. Without classification cost, all detectors significantly drop the detection accuracy and heavily rely on NMS. Instead, adding classification cost eliminates the necessity of NMS.

3.3. Matching Cost

Location cost. By reviewing non-end-to-end object detectors, we identity that they assign positive samples by only location cost. The location cost is defined as follows:

$$C_{loc} = \lambda_{iou} \cdot C_{iou} + \lambda_{L1} \cdot C_{L1} \tag{4}$$

where C_{L1} and C_{iou} are L1 loss and IoU loss between sam-

ple and ground-truth box, respectively. λ_{L1} and λ_{iou} are coefficients. When object candidates are points in the feature map, $\lambda_{iou} = 0$. We note that object candidates could be pre-defined or predicted. Take an example of RetinaNet, its pre-defined object candidates are anchor boxes, while its predicted object candidates are predicted boxes refined by the predicted offsets. As for CenterNet and FCOS, the pre-defined object candidates are grid points in the feature map, while the predicted object candidates are predicted boxes. Based on object candidates, the location cost could also be pre-defined or predicted.

Location cost can reasonably measure whether the selected positive sample is beneficial for location. However, object detection is a multi-task of location and classification. Classification cost is supposed to be considered as well, although it has been ignored for decades before recently.

Classification cost. By introducing classification cost into assignment, the total cost is the summation of classification cost and location cost between sample and ground-truth, defined as follows:

$$C = \lambda_{cls} \cdot C_{cls} + C_{loc} \tag{5}$$

where C_{cls} is classification loss of predicted classifications and ground truth category labels. C_{loc} is defined in Equation 4. λ_{cls} is coefficient.

As shown in Table 2, the default settings of end-to-end object detectors include both location cost and classification cost. When discarding classification cost, these detectors significantly degenerate and heavily rely on NMS.

Continuing on one-to-one assignment versions of RetinaNet, CenterNet and FCOS, classification cost is additionally introduced to their matching cost. For RetinaNet and CenterNet, the positive sample is selected as the sample with minimum matching cost among all samples. For FCOS, the positive sample is chosen from the pre-defined layer in feature pyramids. As shown in Table 2, adding classification cost immediately makes NMS has little effect on the detection performance.

Conclusion 3.2 *Non-end-to-end detectors assign positive samples by only location cost. However, when additionally considering classification cost, they immediately produce one-to-one prediction under one-to-one assignment.*

To completely reduce the necessity of NMS, we also carry out the experiments in which the pre-defined location cost in RetinaNet, CenterNet and FCOS is changed to predicted location cost. We note that the location cost in DETR, Deformable DETR and Sparse R-CNN is also based on predicted boxes. As shown in Table 2, the combination of classification cost and predicted location cost enable previous non-end-to-end detectors to achieve completely end-to-end. Interestingly, the location cost based on predicted boxes could obtain better detection performance. We explain it is because predicted location cost makes matching cost more aligned to the training loss function, thus benefits to the optimization of the object detector.

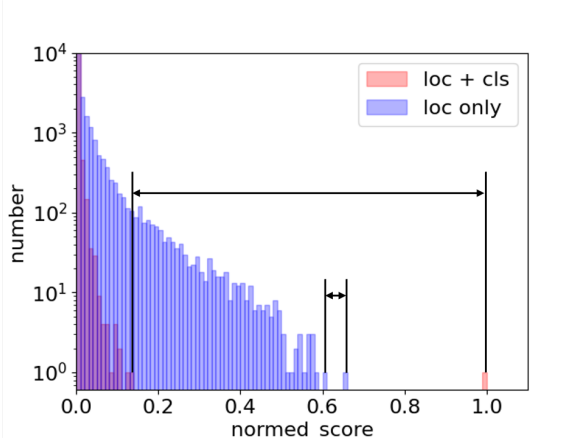


Figure 4. **Samples' classification scores of the trained detector.** For better visualization, we only show the part below number of 10^4, and scores are normalized to [0, 1]. Blue bins show the detector trained with positive samples chosen by only location cost. Red bins consider both location cost and classification cost. Classification cost results in a clear score gap between samples of first highest score and second highest score.

Our experiments above demonstrate that one-to-one assignment is necessary but not sufficient for one-to-one prediction. Additionally considering classification cost is the key to achieve end-to-end object detection. We further explore how classification cost makes an effect.

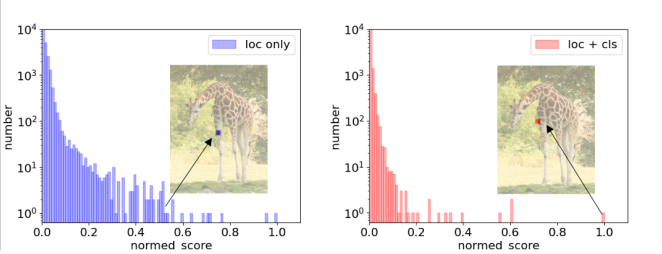
3.4. Score Gap

In order to understand how classification cost contributes to end-to-end object detection, we first introduce the following definition.

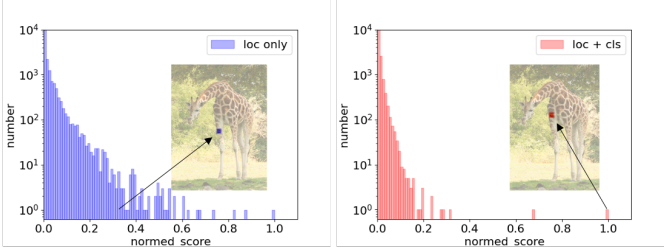
Definition 3.3 (Score Gap) *Given a classification network $\mathcal{N}$ and a set of samples $\mathcal{S}$, if sample i is positive and others are negative, train the network and get each sample's score s(i), let $i_{max} = \arg\max_{j \in \mathcal{S}} s(j)$, then score gap ($\mathcal{N}$, $\mathcal{S}$, i) is defined as:*

$$\text{score gap}(\mathcal{N}, \mathcal{S}, i) = \min_{j \in \mathcal{S} \setminus i_{max}} (s(i_{max}) - s(j)) \tag{6}$$

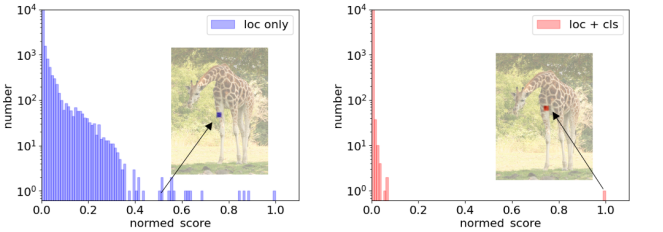
The score gap describes the gap between the first-highest score and the second-highest score. A sufficient requirement for end-to-end object detection is that the score gap



(a) Training early stage.



(b) Training middle stage.



(c) Training late stage.

Figure 5. **Positive samples in different training stages.** For better visualization, we only show the part below the number of 10^4, and scores are normalized to [0, 1]. Blue bins show the detector trained with positive samples chosen by only location cost. Red bins consider both location cost and classification cost. Only location cost selects the positive samples as those of medium score. Introducing classification loss makes positive samples as those of the highest score during the whole training process.

should be large enough, otherwise, non-maximum predictions cannot be easily filtered out: a high score threshold may filter out all predictions, while a low threshold may output redundant predictions.

In Figure 4, we show samples' classification scores from the trained detectors with and without classification cost under one-to-one assignment. For only location cost, the gap between the highest score and the second-highest score is negligible. Also, all samples are relatively lower scores. Instead, considering classification cost produces a clear score gap, therefore, achieves end-to-end object detection.

To explore how score gap is produced under different matching costs, we further show samples' classification scores during different training stages in Figure 5.

For only location cost, the positive sample lies in the grid point closest to the center of the object ground-truth box. Nevertheless, the positive samples are those of medium score. Such positive samples will push the network to pull down the score of samples that have been high score. As a consequence, all samples tend to be relatively lower scores.

When additionally considering classification cost, the positive sample is those of highest score in the training iterations. These choices are much more useful to further increase the score of positive samples and widen the score gap, meanwhile, it does not hurt the box regression since the positive samples are still inside the object ground-truth box. After the whole training process, a large enough score gap is finally generated to achieve end-to-end object detection.

Conclusion 3.4 *Classification cost chooses positive samples as those of highest score in the training process, therefore, produces large enough score gap for end-to-end object detection.*

We note that the positive sample selected by only location cost is the same sample during the whole training process, but classification score of this sample is always kept as the medium score. To explain why its score can't be lifted, we visualize positive samples in different training images, as shown in Figure 6.

If only considering location cost, the positive sample lies in the grid point closest to the center of the object ground-truth box. This assignment is beneficial for box regression, but is not a good choice for foreground and background classification. Specifically, some background-like samples are assigned as positive samples, highlighted by yellow rings in Figure 6. These cases come from objects' arbitrary shapes and poses, such as the long neck of the giraffe. These background-like samples are noisy samples for classification task and decrease the discriminative ability of the network.

On the contrary, when classification cost is introduced, positive samples are grid points in more discriminative areas, *e.g.*, neck of the giraffe. In such cases, it is avoided to select positive samples outside the area of the object. Moreover, these discriminative positive samples are also more useful for classification branch to distinguish noisy samples. As a result, noisy samples are effectively reduced from positive samples.

Observation 3.5 *Only location cost may select noisy background-like positive samples, while additionally considering classification cost could reduce these noisy positive samples.*

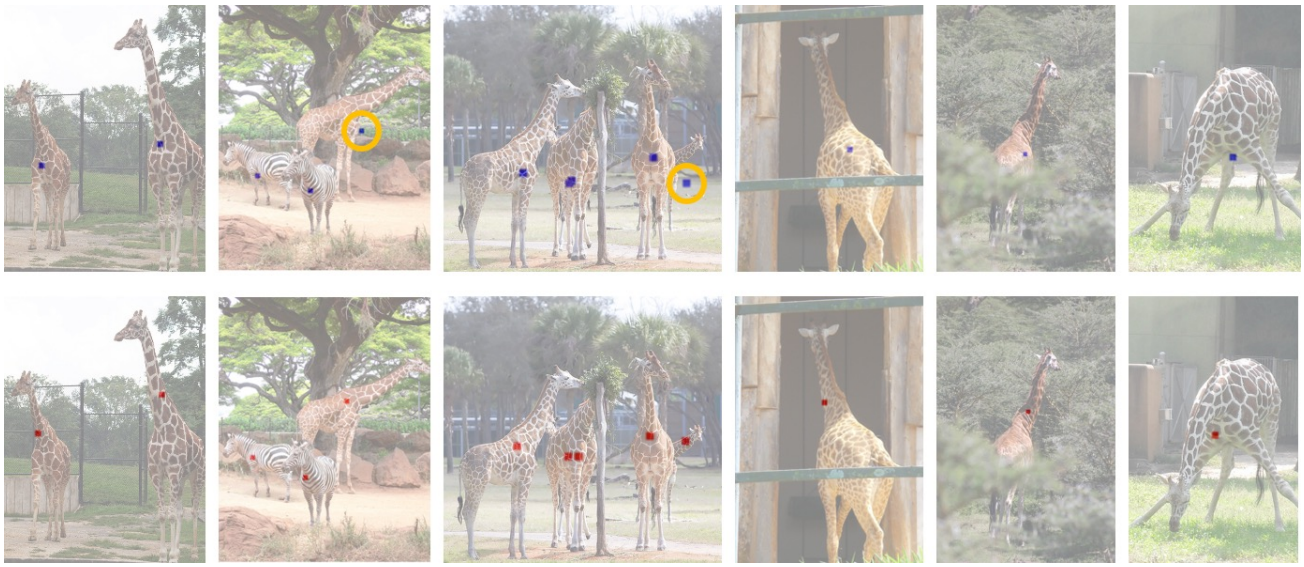


Figure 6. **Positive samples in different training images.** For better visualization, the positive grid points are highlighted by surrounding circles. 1st row is only location cost. 2nd row is the summation of classification cost and location cost. The positive samples assigned by only location cost are the grid point closest to the ground-truth box center, however, some background-like samples are assigned as positive samples, highlighted by yellow rings. Adding classification cost, positive samples are grid points in more discriminative areas, *e.g.*, neck of the giraffe.

From the above analysis, we discover that non-end-to-end detectors only consider location cost to select positive samples, which makes noisy positive samples and decreases the discriminative ability of the network. This leads to a small score gap and produces redundancy predictions. Instead, when classification cost is additionally introduced, the noisy samples could be reduced, the score gap is large enough, therefore, end-to-end object detection is achieved.

4. Theoretical Analysis

4.1. Setup

In this section, we analyze the convergence properties of object detectors under one-to-one assignment with matching cost as the summation of location cost and classification cost, in which only one sample with the minimum matching cost is assigned as the positive sample, others are all negative samples.

Since a systematic framework is beyond our reach, we first make some reasonable assumptions based on verification experiments. We conduct the experiment in which the parameter of location sub-net is fixed, only classification sub-net is trained. And we observe the same conclusions as Section 3. This leads to the following observation:

Observation 4.1 *The optimization of classification sub-net is irrelevant to location sub-net.*

Based on Observation 4.1, analysis of classification score of object detection can be reasonably simplified as a single classification problem, in which only the sample with minimum classification cost is chosen as the positive sample among all samples, others are all negative samples.

We focus on analyzing properties using linear classifier. Let $\mathcal{X} = \{x \in \mathbb{R}^d : \|x\| \leq 1\}$ be an instance space and $\mathcal{Y} = \{+1, -1\}$ be the label space. The label of a positive sample is $+1$ while the label of a negative sample is -1. We wish to train a classifier h, coming from a hypothesis class $\mathcal{H} = \{x \mapsto \text{sign}(w^T x) : w \in \mathbb{R}^d\}$. Note that we can express the bias term b by rewriting $w = [\hat{w}, b]$ and $x = [\hat{x}, 1]$. We use the perceptron's update rule with mini-batch size of 1. That is, given the classifier $w_t \in \mathbb{R}^d$, the update is only performed on incorrectly classified example $(x_t, y_t) \in \mathcal{X} \times \mathcal{Y}$ as given by $w_{t+1} = w_t + \eta y_t x_t$ where η is the stepsize. According to one-to-one positive label assignment, in each update step, we denote $x_t^1 = \arg\max_{x \in \mathcal{X}} {w_t}^{\mathsf{T}} x$, the label of x_t^1 is $y(x_t^1) = +1$ and the labels of the remaining samples in $\mathcal{X}$ are $y(x) = -1, x \in \mathcal{X} \backslash \{x_t^1\}$.

4.2. Theoretical Results

We first show that samples with labels assigned by one-to-one assignment are linearly separable at each training iteration, implying the positive definite score gap. Based on this result, then we show that the one-to-one assignment can converge within finite update steps.

Proposition 4.2 (Feasibility) *Suppose that the one-to-one assignment is run on a sequence of examples from $\mathcal{X} \times \mathcal{Y}$. Given weight vector $w_t = [\hat{w}_t, b_t]$ at update step t, there exists $\gamma_t \in \mathbb{R}$ and $\delta_t > 0$ such that for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ we have $y(w_t^*)^T x \geq \delta_t$ with $w_t^* = [\hat{w}_t, \gamma_t]$.*

Proof. Detailed proof is provided in Appendix.

By Proposition 4.2, we see that there always exists a classifier that can correctly classify all samples at every update step when the label is assigned by one-to-one assignment.

Theorem 4.3 (Convergence) *Let γ_{t+1} and γ_t be the constants defined in Proposition 4.2. For each update step t, we assume there exists a stepsize η_t such that $\|x_t\|^2 \eta_t^2 + y_t(\gamma_{t+1} - 2\gamma_t)\eta_t + b_t(\gamma_{t+1} - \gamma_t) > 0$ where (x_t, y_t) be the incorrectly classified sample at iteration t. If the sample label is assigned by one-to-one assignment, then, $t \leq \frac{\eta_{max}^2 - 2\eta_{min}\delta_{\min}(w_1^{\mathsf{T}} w_0^* - \|w_0\| - \eta_{max})}{2\eta_{min}^2 \delta_{\min}^2}$ where η_{max} and η_{min} are the maximum and the minimum value of stepsize among all t's updates, w_1 is the classifier after the first update and $\delta_{\min}$ is the minimum of all δ_ts in Proposition 4.2. All instances at initialization can be correctly classified by w_0^*.*

Proof. Detailed proof is provided in Appendix.

Theorem 4.3 shows that samples with labels assigned by one-to-one assignment can converge to a classifier that admits a single positive sample, *i.e.*, its label is $+1$. Therefore it is guaranteed to converge to a solution in the sense that the classification output is one-to-one prediction.

Remark 4.4 *The one-to-one assignment with classification cost is guaranteed to converge to a solution in the sense that the classification output is one-to-one prediction. But assignment with only location cost may produce multiple positive samples.*

By Theorem 4.3, we see the one-to-one prediction is based on that there exists a classifier that can correctly classify all samples at every update step. However, without classification cost, only location cost determines positive samples by location criterion, which may cause the problem that such the positive sample may not be linearly separable with the remaining negative samples. In this case, perception learning algorithm can converge to a classifier that makes the fewest errors in prediction (Burton et al., 1997). Hence, it is likely to produce many positive samples.

5. Crowded Object Detection

In crowded scenarios, previous non-end-to-end detectors suffer from one dilemma when using NMS to remove duplicate predictions (Zhang et al., 2019): higher NMS threshold brings more false positives, while a lower threshold may mistakenly remove true positives and cause undetected objects. On the contrary, end-to-end detectors completely avoid this problem by eliminating NMS, and exhibit superior performance in In crowded scenes.

Method	NMS	AP_{50}	mMR↓	Recall
Annotation boxes	✓	-	-	95.0
RetinaNet (Lin et al., 2017b)	✓	81.7	57.6	88.6
RetinaNet + o2o + cls.	○	90.8	49.3	98.1
	✓	86.3	49.9	93.2
FCOS (Tian et al., 2019)	✓	86.1	55.2	94.3
FCOS + o2o + cls.	○	90.7	48.2	97.6
	✓	86.0	49.0	92.3

Table 3. **Comparisons with different object detectors on CrowdHuman validation set.** "○" means no NMS processing. Annotation boxes processed by NMS only obtain 95.0% recall, which is the upper bound of non-end-to-end detectors. End-to-end versions of RetinaNet and FCOS are not constrained to that recall upper bound and outperform their baselines setting by a large margin. NMS damages the performance of end-to-end detectors on crowded scenes.

5.1. Experiment Setting

Detectors. We select RetinaNet (Lin et al., 2017b), FCOS (Tian et al., 2019) and their end-to-end variants with predicted location cost.

Dataset. CrowdHuman (Shao et al., 2018) is a widely-used benchmark for crowded object detection, in which human boxes are highly crowded and overlapped. We use metrics AP, mMR and recall of on IoU 0.5 threshold. All models are trained on training set (∼15k images) and evaluated with validation set (∼4k images).

5.2. Results

Table 3 shows the performance of different object detectors on CrowdHuman. We first show that applying NMS on annotation boxes only obtains 95% recall, which indicates the even strongest non-end-to-end detectors are bounded to NMS in crowded scenes. Instead, when we reform RetinaNet (Lin et al., 2017b) and FCOS (Tian et al., 2019) to end-to-end detectors by adding one-to-one positive sample and classification cost, they are not constrained to this recall upper bound and significantly improve the recall to 98.1% and 97.6%, respectively. Meanwhile, AP_{50} and mMR benefit a large improvement from end-to-end setting. When NMS is used to process the predictions of end-to-end RetinaNet and FCOS, the performance degenerates at once. It furthermore demonstrates the disadvantage of NMS in crowded scenes and the superiority of end-to-end object detection.

6. Related Work

Object detection. Object detection is one of the most fundamental and challenging topics in computer vision fields. Limited by classical feature extraction techniques (Dalal & Triggs, 2005; Viola & Jones, 2001), the performance has plateaued for decades, and the application scenarios are limited. With the rapid development of deep learning (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015; Szegedy et al., 2015; He et al., 2016; Huang et al., 2017), object detection achieves powerful performance (Everingham et al., 2010; Lin et al., 2014).

One-stage detector. One-stage detector directly predicts the category and location of dense anchor boxes or points over different spatial positions and scales in a single-shot manner such as YOLO (Redmon et al., 2016), SSD (Liu et al., 2016) and RetinaNet (Lin et al., 2017b). YOLO (Redmon et al., 2016) divides the image into an S × S grid, and if the center of an object falls into a grid cell, the corresponding cell is responsible for detecting this object. SSD (Liu et al., 2016) directly predicts object category and anchor box offsets on multi-scale feature map layers. RetinaNet (Lin et al., 2017b) utilizes focal loss to ease the extreme unbalance of positive and negative samples based on the FPN (Lin et al., 2017a). Recently, anchor-free detectors (Huang et al., 2015) is proposed to make this pipeline much simpler by replacing hand-crafted anchor boxes with reference points. CornerNet (Law & Deng, 2018) generates the keypoints by heatmap and group them by the Associative Embedding (Newell et al., 2017). CenterNet (Zhou et al., 2019) directly uses the center point to regress the target object on a single scale. FCOS (Tian et al., 2019) assigns the objects of different size and scales to multi-scale feature maps with the power of FPN (Lin et al., 2017a). ATSS(Zhang et al., 2020b) reveals that the essential difference between anchor-based and anchor-free detection is how to define positive and negative training samples, leading to the performance gap between them.

Two-stage detector. The two-stage detectors (Cai & Vasconcelos, 2018; Dai et al., 2016; Girshick, 2015; He et al., 2017; Ren et al., 2015) firstly generate a high-quality set of foreground proposals by region proposal networks and then refine each proposal's location and predicts its category. Fast R-CNN (Girshick, 2015) uses Selective Search (Uijlings et al., 2013) to generate foreground proposals and refine the proposals in R-CNN (Girshick et al., 2014) Head. Faster R-CNN (Ren et al., 2015) proposes the region proposal network, which generates high-quality proposals in real-time. Cascade R-CNN (Cai & Vasconcelos, 2018) iteratively uses multiple R-CNN heads with different label assign threshold to get high-quality detection boxes. Cascade RPN (Vu et al., 2019) improves the region proposal quality and detection performance by systematically addressing the limitation of the conventional RPN that heuristically defines the anchors and aligns the features to the anchors. Libra R-CNN (Pang et al., 2019) tries to solve the unbalance problems in sample level, feature level, and objective level. Grid R-CNN (Lu et al., 2019) adopts a grid-guided localization mechanism for accurate object detection instead of traditional bounding box regression.

End-to-end object detection. The well-established end-to-end object detectors are based on sparse candidates and multiple-stage refinement. Relation Network (Hu et al., 2018) and DETR (Carion et al., 2020) directly output the predictions without any hand-crafted assignment and post-processing procedure, achieving fantastic performance. DETR utilizes a sparse set of object queries to interact with the global image feature. Benefit from the global attention mechanism (Vaswani et al., 2017) and the bipartite matching between predictions and ground truth objects, DETR can discard the NMS procedure while achieving remarkable performance. Deformable-DETR (Zhu et al., 2020) is introduced to restrict each object query to a small set of crucial sampling points around the reference points, instead of all points in the feature map. Sparse R-CNN (Sun et al., 2020a) starts from a fixed sparse set of learned object proposals and iteratively performs classification and localization to the object recognition head. Adaptive Clustering Transformer (Zheng et al., 2020) proposes to improve the attention in DETR's encoder by LSH approximate clustering. UP-DETR (Dai et al., 2020) improves the convergence speed of DETR by a self-supervised method. TSP (Sun et al., 2020b) analyzes co-attention and bipartite matching are two main causes of slow convergence in DETR. SMCA (Gao et al., 2021) explores global information with a self-attention and co-attention mechanism to achieve fast convergence and better accuracy performance.

7. Conclusion

Assigning positive samples by location cost is conceptually intuitive and popularizes in object detection to date. However, in this work, we surprisingly find that this widely-used method is the obstacle of end-to-end detectors. By additionally considering classification cost, previous detectors immediately achieve end-to-end detection. Our findings uncover that answer to the notorious problem of defining positive samples in object detection is embarrassingly simple: in every training iteration, selecting only one positive sample which could minimize training loss is just 'right'.

Acknowledgements

This work was supported by the General Research Fund of HK No.27208720.

References

Burton, R. M., Herold G, D., and Rienk S, V. Perceptron algorithms for the classification of non-separable populations. *Stochastic Models*, 13(2):205–222, 1997. 8

Cai, Z. and Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In *CVPR*, 2018. 1, 2, 3, 9

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-End object detection with transformers. In *ECCV*, 2020. 1, 2, 3, 9

Dai, J., Li, Y., He, K., and Sun, J. R-FCN: Object detection via region-based fully convolutional networks. In *NeurIPS*, 2016. 9

Dai, Z., Cai, B., Lin, Y., and Chen, J. Up-detr: Unsupervised pre-training for object detection with transformers. *arXiv preprint arXiv:2011.09094*, 2020. 9

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 1, 9

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., and Zisserman, A. The pascal visual object classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010. 9

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. Object detection with discriminatively trained part based models. *T-PAMI*, 32(9):1627–1645, 2010. 1

Gao, P., Zheng, M., Wang, X., Dai, J., and Li, H. Fast convergence of detr with spatially modulated co-attention. *arXiv preprint arXiv:2101.07448*, 2021. 9

Girshick, R. Fast R-CNN. In *ICCV*, 2015. 1, 9

Girshick, R., Donahue, J., Darrell, T., and Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 1, 2, 3, 9

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016. 9

He, K., Gkioxari, G., Dollar, P., and Girshick, R. Mask R-CNN. In *ICCV*, 2017. 9

Hu, H., Gu, J., Zhang, Z., Dai, J., and Wei, Y. Relation networks for object detection. In *CVPR*, 2018. 2, 9

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *CVPR*, 2017. 9

Huang, L., Yang, Y., Deng, Y., and Yu, Y. DenseBox: Unifying landmark localization with end to end object detection. *arXiv preprint arXiv:1509.04874*, 2015. 9

Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 9

Kuhn, H. W. The Hungarian method for the assignment problem. *NRL*, 2(1-2):83–97, 1955. 4

Law, H. and Deng, J. CornerNet: Detecting objects as paired keypoints. In *ECCV*, 2018. 9

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 4, 9

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. Feature pyramid networks for object detection. In *CVPR*, 2017a. 9

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollar, P. Focal loss for dense object detection. In *ICCV*, 2017b. 1, 2, 3, 8, 9

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., and Berg, A. C. SSD: Single shot multibox detector. In *ECCV*, 2016. 1, 9

Lu, X., Li, B., Yue, Y., Li, Q., and Yan, J. Grid R-CNN. In *CVPR*, 2019. 9

Newell, A., Huang, Z., and Deng, J. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in neural information processing systems*, pp. 2277–2287, 2017. 9

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., and Lin, D. Libra R-CNN: Towards balanced learning for object detection. In *CVPR*, 2019. 9

Redmon, J. and Farhadi, A. YOLO9000: Better, faster, stronger. In *CVPR*, 2017. 1

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 1, 2, 9

Ren, S., He, K., Girshick, R., and Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 1, 2, 9

Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., and Savarese, S. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 2

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., and Sun, J. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 2, 8

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 9

Sun, P., Zhang, R., Jiang, Y., Kong, T., Xu, C., Zhan, W., Tomizuka, M., Li, L., Yuan, Z., Wang, C., et al. Sparse r-cnn: End-to-end object detection with learnable proposals. *arXiv preprint arXiv:2011.12450*, 2020a. 2, 3, 9

Sun, Z., Cao, S., Yang, Y., and Kitani, K. Rethinking transformer-based set prediction for object detection. *arXiv preprint arXiv:2011.10881*, 2020b. 9

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *CVPR*, 2015. 9

Tian, Z., Shen, C., Chen, H., and He, T. FCOS: Fully convolutional one-stage object detection. In *ICCV*, 2019. 1, 2, 3, 4, 8, 9

Uijlings, J. R., Van De Sande, K. E., Gevers, T., and Smeulders, A. W. Selective search for object recognition. *IJCV*, 104(2):154–171, 2013. 9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017. 9

Viola, P. and Jones, M. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, pp. I–I. IEEE, 2001. 1, 9

Vu, T., Jang, H., Pham, T. X., and Yoo, C. D. Cascade RPN: Delving into high-quality region proposal network with adaptive convolution. In *NeurIPS*, 2019. 9

Zhang, H., Chang, H., Ma, B., Wang, N., and Chen, X. Dynamic R-CNN: Towards high quality object detection via dynamic training. In *ECCV*, 2020a. 3

Zhang, K., Xiong, F., Sun, P., Hu, L., Li, B., and Yu, G. Double anchor r-cnn for human detection in a crowd. *arXiv preprint arXiv:1909.09998*, 2019. 2, 3, 8

Zhang, S., Chi, C., Yao, Y., Lei, Z., and Li, S. Z. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *CVPR*, 2020b. 3, 9

Zheng, M., Gao, P., Wang, X., Li, H., and Dong, H. End-to-end object detection with adaptive clustering transformer. *arXiv preprint arXiv:2011.09315*, 2020. 9

Zhou, X., Wang, D., and Krähenbühl, P. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 1, 2, 3, 9

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2, 3, 9