Co-DETR

DETRs with Collaborative Hybrid Assignments Training

2025.02.20

Seung min chung

Contents

01 Abstract

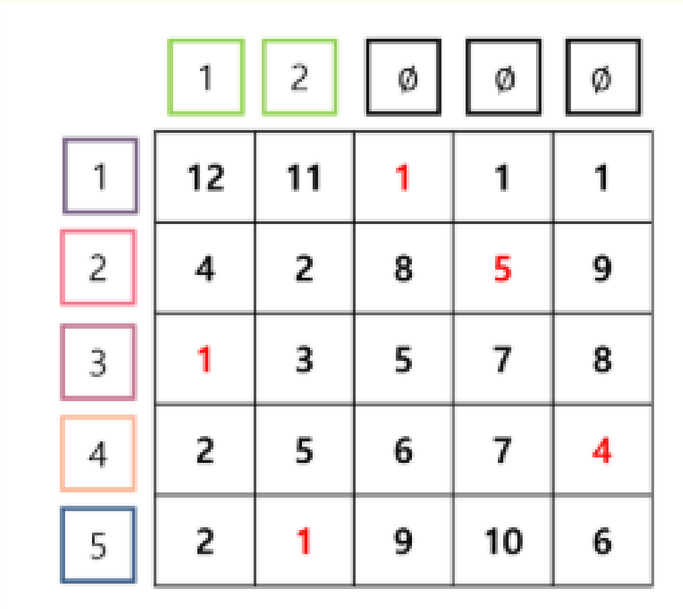
Main Idea

DETR relies on a limited number of positive queries, so they can’t train efficiently. In this paper they alleviate this problem incorporating an auxiliary head in the decoder layer. Auxiliary head make new positive queries, leading to improve training efficiently

Core

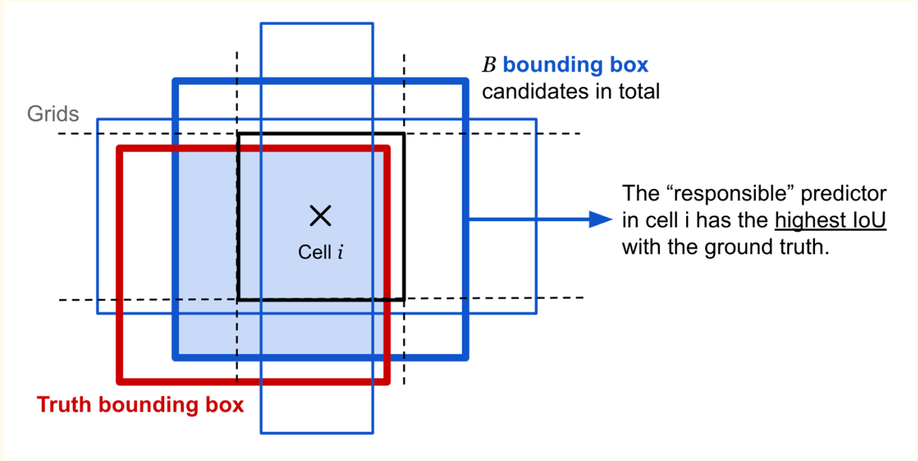
- Incorporating an auxiliary head
- 1 : N matching

02 Introduction



DETR

DETR uses only N queries and employes on—to-one matching. However, since there are only a few object in a given image, most of the queries turn out to be negative. This leads to inefficient training



Co-DETR

In object-detection task, one-to-many matching is a well-known technique to improve accuracy.

Co-DETR increases the number of positive queries by incorporating an one-to-many matching auxiliary head in the decoder layer.

02 Introduction

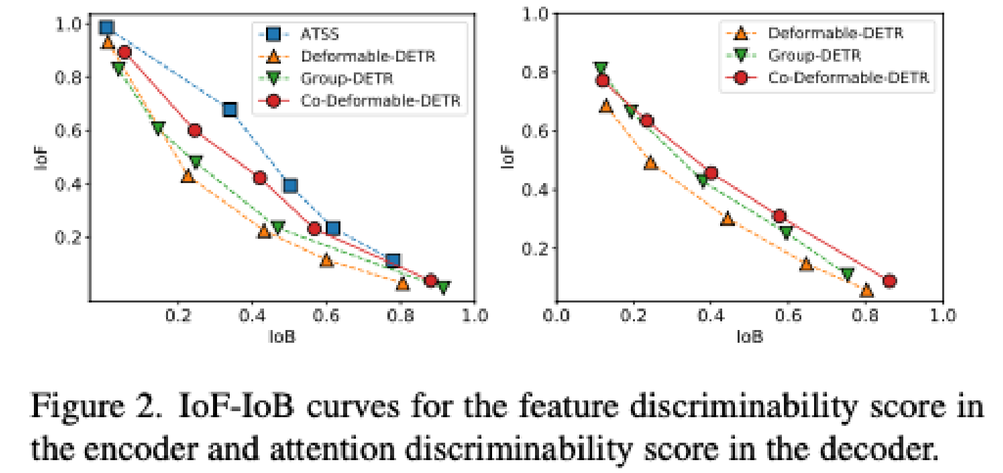


Figure 2. IoF-IoB curves for the feature discriminability score in the encoder and attention discriminability score in the decoder.

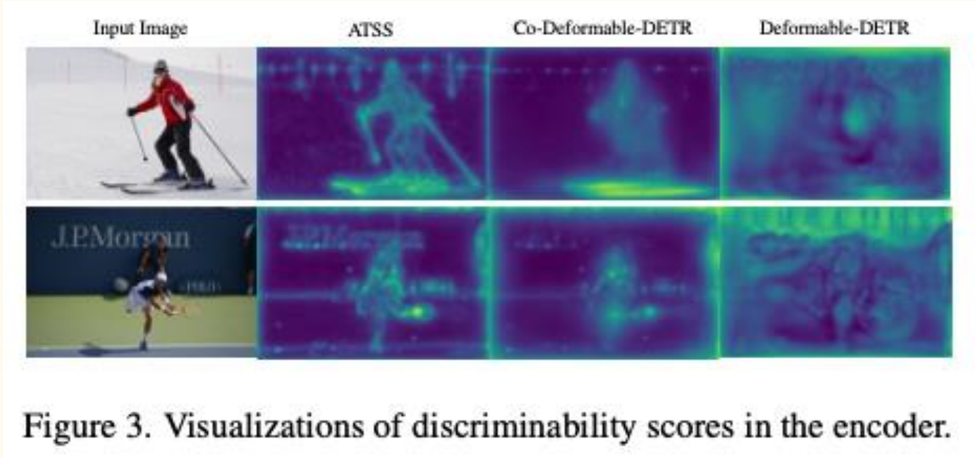


Figure 3. Visualizations of discriminability scores in the encoder.

Object-detection task is composed of localization and classification. Traditional detection models typically employe one-to-many matching, such as R-CNN, ATSS and perform final matching using NMS. These one-to-many matching techniques are highly effective in object- detection.

Figure2 shows that one-to-many matching methods, such as ATSS have a higher discriminability score in the encoder compared to DETR. This means that one-to-many matching can learn more diverse patterns in feature map.

Figure3 shows a visualization of the encoder. This visualization indicates that one-to-many method can detect details more attentively.

03 collaborative Hybrid Assignments Training

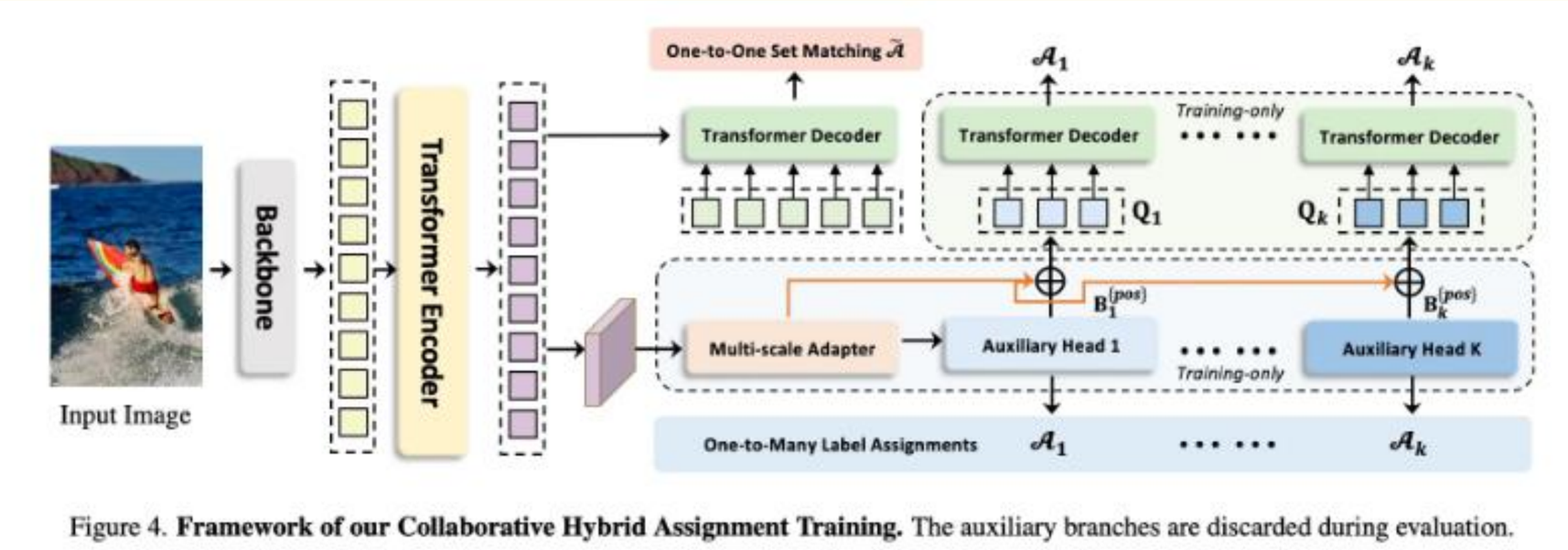


Figure 4. **Framework of our Collaborative Hybrid Assignment Training.** The auxiliary branches are discarded during evaluation.

Co-DETR architecture have 2 stream.

(a) Encoder – Decoder (one-to-one matching) : original DETR stream

(b) Encoder – Multi-scale Adapter – Auxiliary Head – Decoder : collaborative Hybrid Assignments Training stream

Multi-Scale Adapter obtain the feature map from encoder and constructs a feature pyramid with J levels. Then K auxiliary head process this J feature pyramid

Each auxiliary head is simple MLP and employes an independent one-to-many matching algorithm such as those used in Faster R-CNN, ATSS. Through this process, each auxiliary head can generate independent positive queries.

04 customized Positive Queries Generation

Head i	Loss $\mathcal{L}_i$	Assignment $\mathcal{A}_i$		
		$\{pos\}$, $\{neg\}$ Generation	$\mathbf{P}_i$ Generation	$\mathbf{B}_i^{\{pos\}}$ Generation
Faster-RCNN [27]	cls: CE loss, reg: GIoU loss	$\{pos\}$: IoU(proposal, gt)>0.5 $\{neg\}$: IoU(proposal, gt)<0.5	$\{pos\}$: gt labels, offset(proposal, gt) $\{neg\}$: gt labels	positive proposals (x_1, y_1, x_2, y_2)
ATSS [41]	cls: Focal loss reg: GIoU, BCE loss	$\{pos\}$:IoU(anchor, gt)>(mean+std) $\{neg\}$: IoU(anchor, gt)<(mean+std)	$\{pos\}$: gt labels, offset(anchor, gt), centerness $\{neg\}$: gt labels	positive anchors (x_1, y_1, x_2, y_2)
RetinaNet [21]	cls: Focal loss reg: GIoU Loss	$\{pos\}$: IoU(anchor, gt)>0.5 $\{neg\}$: IoU(anchor, gt)<0.4	$\{pos\}$: gt labels, offset(anchor, gt) $\{neg\}$: gt labels	positive anchors (x_1, y_1, x_2, y_2)
FCOS [32]	cls: Focal Loss reg: GIoU, BCE loss	$\{pos\}$: points inside gt center area $\{neg\}$: points outside gt center area	$\{pos\}$: gt labels, ltrb distance, centerness $\{neg\}$: gt labels	FCOS point (cx, cy) $w = h = 8 \times 2^{2+j}$

Table 1. **Detailed information of auxiliary heads.** The auxiliary heads include Faster-RCNN [27], ATSS [41], RetinaNet [21], and FCOS [32]. If not otherwise specified, we follow the original implementations, *e.g.*, anchor generation.

Until now, we have generated additional positive queries to address scarcity of positive queries.

$$Q_i = \text{Linear}(PE(B_i^{pos})) + \text{Linear}(E(F_*, pos))$$

Now we have K + 1 group of queries : one group consists of the original one-to-one queries, and the remaining k groups are generated by the auxiliary heads. Since the K groups of queries are positive, no matching algorithm is applied.

$$\mathcal{L}^{global} = \sum_{l=1}^{L} \left(\tilde{\mathcal{L}}l^{dec} + \lambda_1 \sum i = 1^K \mathcal{L}_{i,l}^{dec} + \lambda_2 \mathcal{L}^{enc} \right)$$

The final loss consists of the original one-to-one loss, one-to-many loss, and encoder loss.

05 Why Co-DETR works ?

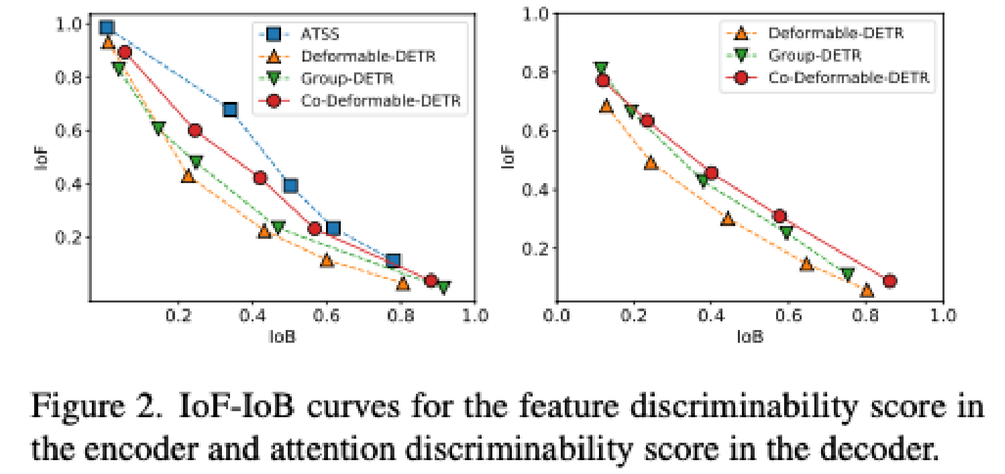


Figure 2. IoF-IoB curves for the feature discriminability score in the encoder and attention discriminability score in the decoder.

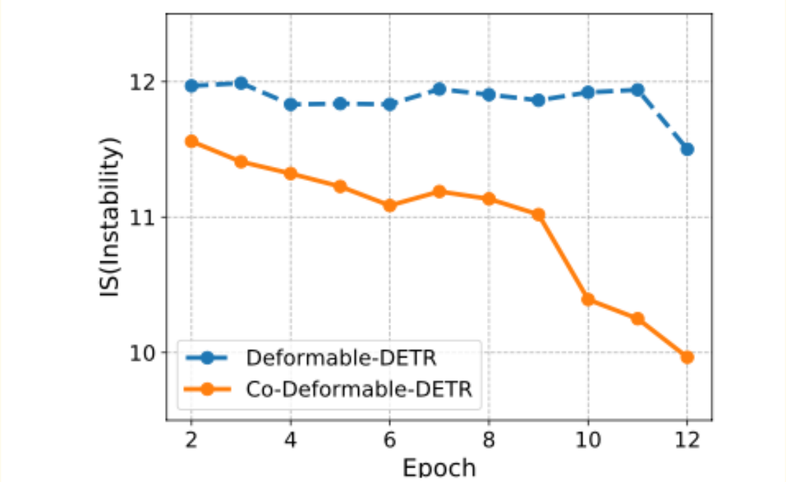


Figure 5. The instability (IS) [18] of Deformable-DETR and *Co*-Deformable-DETR on COCO dataset. These detectors are trained for 12 epochs with ResNet-50 backbones.

Enrich the encoder's supervisions

According to Figure2, using one-to-many method leads to more effective training in the encoder. Consequently, one-to-many method shows a higher discriminability score in encoder layer. This indicates that the one-to-many method induces the model to learn more diverse patterns in images

Reducing the instability

During training, the result of the Hungarian matching algorithm changes for each epoch.
Consequently, this causes instability in training. However Co-DETR discards the Hungarian matching, making it more stable than DETR

06 Experiments

Performance improvement when transforming DETR into Co-DETR

Method	K	#epochs	AP
Conditional DETR-C5 [26]	0	36	39.4
Conditional DETR-C5 [26]	1	36	41.5(**+2.1**)
Conditional DETR-C5 [26]	2	36	41.8(**+2.4**)
DAB-DETR-C5 [23]	0	36	41.2
DAB-DETR-C5 [23]	1	36	43.1(**+1.9**)
DAB-DETR-C5 [23]	2	36	43.5(**+2.3**)
Deformable-DETR [43]	0	12	37.1
Deformable-DETR [43]	1	12	42.3(**+5.2**)
Deformable-DETR [43]	2	12	42.9(**+5.8**)
Deformable-DETR [43]	0	36	43.3
Deformable-DETR [43]	1	36	46.8(**+3.5**)
Deformable-DETR [43]	2	36	46.5(**+3.2**)

Table 2. Results of plain baselines on COCO `val`.

Performance improvement when adapt Co-DETR in SOTA model

Method	K	#epochs	AP
Deformable-DETR++ [43]	0	12	47.1
Deformable-DETR++ [43]	1	12	48.7(**+1.6**)
Deformable-DETR++ [43]	2	12	49.5(**+2.4**)
DINO-Deformable-DETR† [39]	0	12	49.4
DINO-Deformable-DETR† [39]	1	12	51.0(**+1.6**)
DINO-Deformable-DETR† [39]	2	12	51.2(**+1.8**)
Deformable-DETR++‡ [43]	0	12	55.2
Deformable-DETR++‡ [43]	1	12	56.4(**+1.2**)
Deformable-DETR++‡ [43]	2	12	56.9(**+1.7**)
DINO-Deformable-DETR†‡ [39]	0	36	58.5
DINO-Deformable-DETR†‡ [39]	1	36	59.3(**+0.8**)
DINO-Deformable-DETR†‡ [39]	2	36	59.5(**+1.0**)

Table 3. Results of strong baselines on COCO `val`. Methods with † use 5 feature levels. ‡ refers to Swin-L backbone.

07 Summary

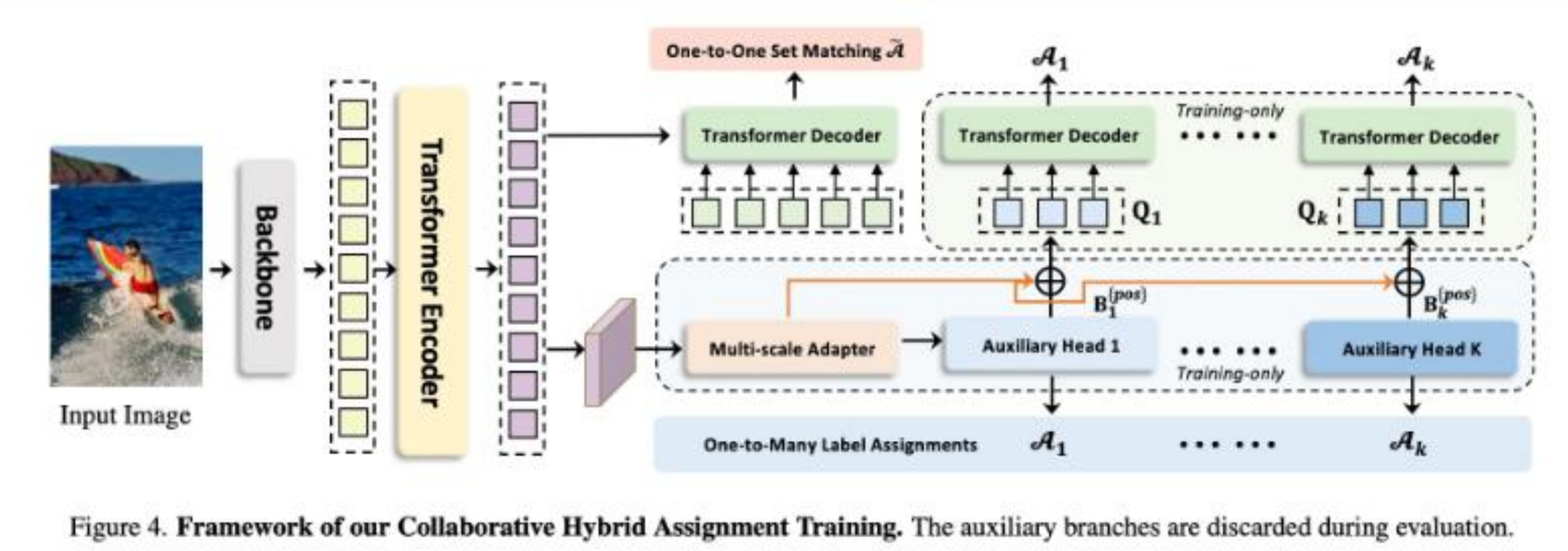


Figure 4. **Framework of our Collaborative Hybrid Assignment Training.** The auxiliary branches are discarded during evaluation.

In this paper, we point out the limited the number of the positive queries in DETR.

To address this problem, Co-DETR incorporates the auxiliary heads in the decoder layer.

Thanks