# DETRs with Collaborative Hybrid Assignments Training

## Supplementary Material

| #convs | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| AP | 41.8 | **42.3** | 41.9 | 42.1 | **42.3** | 42.0 |

Table 1: Influence of number of convolutions in auxiliary head.

| $\lambda_1$ | $\lambda_2$ | #epochs | AP | $AP_S$ | $AP_M$ | $AP_L$ |
|---|---|---|---|---|---|---|
| 0.25 | 2.0 | 36 | 46.2 | 28.3 | 49.7 | 60.4 |
| 0.5 | 2.0 | 36 | 46.6 | 29.0 | 50.5 | 61.2 |
| 1.0 | 2.0 | 36 | **46.8** | **28.1** | **50.6** | **61.3** |
| 2.0 | 2.0 | 36 | 46.1 | 27.4 | 49.7 | 61.4 |
| 1.0 | 1.0 | 36 | 46.1 | 27.9 | 49.7 | 60.9 |
| 1.0 | 2.0 | 36 | **46.8** | **28.1** | **50.6** | **61.3** |
| 1.0 | 3.0 | 36 | 46.5 | 29.3 | 50.4 | 61.4 |
| 1.0 | 4.0 | 36 | 46.3 | 29.0 | 50.1 | 61.0 |

Table 2: Results of hyper-parameter tuning for $\lambda_1$ and $\lambda_2$.

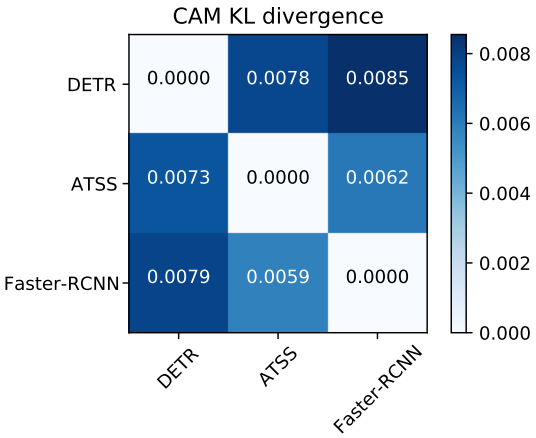


Figure 1: The relation matrix for the DETR head, ATSS head, and Faster-RCNN head. The detector is $\mathcal{C}$o-Deformable-DETR ($K = 2$) with ResNet-50.

## A. More ablation studies

**The number of stacked convolutions.** Table 1 reveals our method is robust for the number of stacked convolutions in the auxiliary head (trained for 12 epochs). Concretely, we simply choose only 1 shared convolution to enable lightweight while achieving higher performance.

**Loss weights of collaborative training.** Experimental results related to weighting the coefficient $\lambda_1$ and $\lambda_2$ are presented in Table 2. We find the proposed method is quite insensitive to the variations of $\{\lambda_1, \lambda_2\}$, since the performance slightly fluctuates when varying the loss coefficients. In summary, the coefficients $\{\lambda_1, \lambda_2\}$ are robust and we set $\{\lambda_1, \lambda_2\}$ to $\{1.0, 2.0\}$ by default.

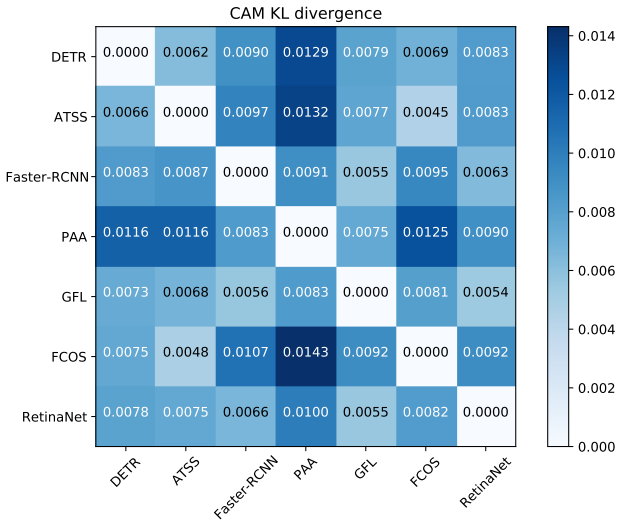


Figure 2: Distances among 7 various heads in our model with $K = 6$.

**The number of customized positive queries.** We compute the average ratio of positive samples in one-to-many label assignment to the ground-truth boxes. For instance, the ratio is 18.7 for Faster-RCNN and 8.8 for ATSS on COCO dataset, indicating more than 8× extra positive queries are introduced when $K = 1$.

**Effectiveness of collaborative one-to-many label assignments.** To verify the effectiveness of our feature learning mechanism, we compare our approach with Group-DETR (3 groups) and $\mathcal{H}$-DETR. First, we find $\mathcal{C}$o-DETR performs better than hybrid matching scheme [2] while training faster and requiring less GPU memory in Table 6. As shown in Table 8, our method ($K = 1$) achieves 46.2% AP, surpassing Group-DETR (44.6% AP) by a large margin even without the customized positive queries generation. More importantly, the IoF-IoB curve in Figure 2 demonstrates Group-DETR fails to enhance the feature representations in the encoder, while our method alleviates the poorly feature learning.

**Conflicts analysis.** We have defined the distance between head $H_i$ and head $H_j$, and the average distance of $H_i$ to measure the optimization conflicts in this study:

$$\mathcal{S}_{i,j} = \frac{1}{|\mathbf{D}|} \sum_{\mathbf{I} \in \mathbf{D}} \mathrm{KL}(\mathcal{C}(H_i(\mathbf{I})), \mathcal{C}(H_j(\mathbf{I})), \tag{1}$$

$$\mathcal{S}_i = \frac{1}{2(K-1)} \sum_{j \neq i}^{K} (\mathcal{S}_{i,j} + \mathcal{S}_{j,i}), \tag{2}$$

where KL, $\mathbf{D}$, $\mathbf{I}$, $\mathcal{C}$ refer to KL divergence, dataset, the input image, and class activation maps (CAM) [8]. In our implementation, we choose the validation set COCO `val` as $\mathbf{D}$ and Grad-CAM as $\mathcal{C}$. We use the output features of DETR encoder to compute the CAM maps. More specifically, we show the detailed distances when $K = 2$ and $K = 6$ in Figure 1 and Figure 2, respetively. The larger distance metric of $\mathcal{S}_{i,j}$ indicates $H_i$ is less consistent to $H_j$ and contributes to the optimization inconsistency.

## B. More implementation details

**One-stage auxiliary heads.** Based on the conventional one-stage detectors, we experiment with various first-stage designs [10, 4, 9, 3, 5] for the auxiliary heads. First, we use the GIoU [7] loss for the one-stage heads. Then, the number of stacked convolutions is reduced from $4$ to $1$. Such modification improves the training efficiency without any accuracy drop. For anchor-free detectors, *e.g.*, FCOS [9], we assign the width of $8 \times 2^j$ and height of $8 \times 2^j$ for the positive coordinates with stride $2^j$.

**Two-stage auxiliary heads.** We adopt the RPN and RCNN as our two-stage auxiliary heads based on the popular Faster-RCNN [6] and Mask-RCNN [1] detectors. To make $\mathcal{C}$o-DETR compatible with various detection heads, we adopt the same multi-scale features (stride 8 to stride 128) as the one-stage paradigm for two-stage auxiliary heads. Moreover, we adopt the GIoU loss for regression in the RCNN stage.

**System-level comparison on COCO.** We first initialize the ViT-L backbone with EVA-02 weights. Then we perform intermediate finetuning on the Objects365 dataset using $\mathcal{C}$o-DINO-Deformable-DETR for 26 epochs and reduce the learning rate by a factor of 0.1 at epoch 24. The initial learning rate is $2.5 \times 10^{-4}$ and the batch size is 224. We choose the maximum size of input images as 1280 and randomly resize the shorter size to $480-1024$. Moreover, we use 1500 object queries and 1000 DN queries for this model. Finally, we finetune $\mathcal{C}$o-DETR on COCO for 12 epochs with an initial learning rate of $5 \times 10^{-5}$ and drop the learning rate at the 8-th epoch by multiplying 0.1. The shorter size of input images is enlarged to $480-1536$ and the longer size is no more than 2400. We employ EMA and train this model with a batch size of 64.

**System-level comparison on LVIS.** In contrast to the COCO setting, we use $\mathcal{C}$o-DINO-Deformable-DETR++ to perform intermediate finetuning on the Objects365 dataset, as we find LSJ augmentation works better on the LVIS dataset. A batch size of 192, an initial learning rate of $2 \times 10^{-4}$, and an input image size of $1280 \times 1280$ are used. We use 900 object queries and 1000 DN queries for this model. During finetuning on LVIS, we arm it with an additional auxiliary mask branch and increase the input size to $1536 \times 1536$. Besides, we train the model without EMA for 16 epochs, where the batch size is set to 64, and the initial learning rate is set to $5 \times 10^{-5}$, which is reduced by a factor of 0.1 at the 9-th and 15-th epoch.

## References


[1] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[2] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Detrs with hybrid matching. *arXiv preprint arXiv:2207.13080*, 2022.

[3] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *European Conference on Computer Vision*, pages 355–371. Springer, 2020.

[4] Xiang Li, Wenhai Wang, Lijun Wu, Shuo Chen, Xiaolin Hu, Jun Li, Jinhui Tang, and Jian Yang. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. *Advances in Neural Information Processing Systems*, 33:21002–21012, 2020.

[5] 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.

[6] 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.

[7] 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.

[8] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[9] 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.

[10] 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.