

Group DETR: Fast DETR Training with Group-Wise One-to-Many Assignment

Qiang Chen^{1*}, Xiaokang Chen^{2*}, Jian Wang¹, Shan Zhang³
Kun Yao¹, Haocheng Feng¹, Junyu Han¹, Errui Ding¹, Gang Zeng², Jingdong Wang^{1†}
¹Baidu VIS

²Key Lab. of Machine Perception (MoE), School of IST, Peking University

³Australian National University

{chenqiang13, wangjian33}@baidu.com

{fenghaocheng, hanjunyu, dingerrui, wangjingdong}@baidu.com

{pkucxk, gang.zeng}@pku.edu.cn, shan.zhang@anu.edu.au

Abstract

Detection transformer (DETR) relies on one-to-one assignment, assigning one ground-truth object to one prediction, for end-to-end detection without NMS post-processing. It is known that one-to-many assignment, assigning one ground-truth object to multiple predictions, succeeds in detection methods such as Faster R-CNN and FCOS. While the naive one-to-many assignment does not work for DETR, and it remains challenging to apply one-to-many assignment for DETR training. In this paper, we introduce Group DETR, a simple yet efficient DETR training approach that introduces a group-wise way for one-to-many assignment. This approach involves using multiple groups of object queries, conducting one-to-one assignment within each group, and performing decoder self-attention separately. It resembles data augmentation with automatically-learned object query augmentation. It is also equivalent to simultaneously training parameter-sharing networks of the same architecture, introducing more supervision and thus improving DETR training. The inference process is the same as DETR trained normally and only needs one group of queries without any architecture modification. Group DETR is versatile and is applicable to various DETR variants. The experiments show that Group DETR significantly speeds up the training convergence and improves the performance of various DETR-based models. Code will be available at <https://github.com/Atten4Vis/GroupDETR>.

1. Introduction

Detection Transformer (DETR) [4] conducts end-to-end object detection without the need for many hand-

*Equal contribution.

†Corresponding author.

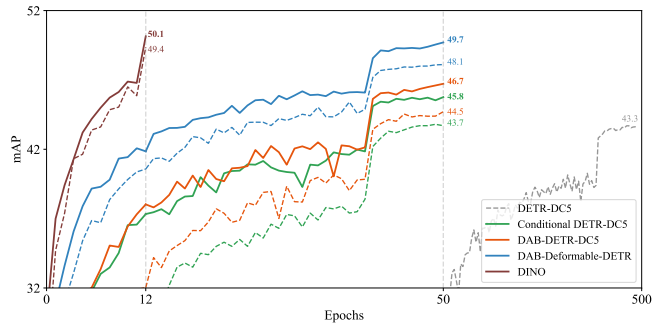


Figure 1. **Group DETR accelerates the training process for DETR variants.** The training convergence curves are obtained on COCO val2017 [34] with ResNet-50 [22]. Dashed and bold curves correspond to the baseline models and the Group DETR counterparts. Best viewed in color.

crafted components, such as non-maximum suppression (NMS) [23] and anchor generation [44, 33, 43]. The architecture consists of a CNN [22] and transformer encoder [53], and a transformer decoder that consists of self-attention, cross-attention and FFNs, followed by class and box prediction FFNs. During training, one-to-one assignment, where one ground-truth object is assigned to one single prediction, is applied for learning to only promote the predictions assigned to ground-truth objects, and demote the duplicate predictions.

This work explores the solutions to accelerate the DETR training process. Previous solutions contain two main lines. The one line is to modify *cross-attention* so that informative image regions are selected for effectively and efficiently collecting the information from image features. Example methods include sparse sampling, through deformable attention [70], and spatial modulations with modifying object queries [16, 41, 8, 57, 61, 36, 17]. The other line is to stabilize *one-to-one assignment* during training, e.g., feeding ground-truth bounding boxes with noises into transformer

decoder [29, 65].

We are interested in the second line. Instead of focusing on stabilizing the assignment like DN-DETR [29], we study the assignment scheme for efficient DETR training from a new perspective: introducing more supervision. It has been proven that assigning one ground-truth object to multiple predictions, i.e., one-to-many assignment, is successful in traditional object detection methods, e.g., Faster R-CNN [44] and FCOS [52] with more anchors and pixels assigned to one ground-truth object. Unfortunately, naive one-to-many assignment does not work for DETR training. It remains a challenge to apply one-to-many assignment to DETR training.

We present a simple yet efficient DETR training approach that uses a group-wise way for one-to-many assignment, called Group DETR. Our approach is based on that end-to-end detection with successful removal of NMS post-processing for DETR comes from the joint effect of two components [4, 41]: decoder self-attention, which collects the information of other predictions, and one-to-one assignment, which expects to learn to score one prediction higher and other duplicate predictions lower for one ground-truth object.

Our approach adopts K groups of object queries, and introduces *group-wise one-to-many assignment*. This assignment scheme conducts one-to-one assignment within each group of object queries, resulting in that one ground-truth object is assigned to multiple predictions. It is encouraged that the prediction assigned to the ground-truth object gets a high score, and other duplicate predictions from the same group of queries get low scores. In other words, the predictions make competition within each group. Thus, our approach uses *separate self-attention*, i.e., self-attention is done for each group separately, eliminating the influence of predictions from other groups and easing DETR training. Regarding inference, it is the same as DETR trained normally, and only needs a single group of object queries.

The resulting architecture is equivalent to DETR with a group of parallel decoders, illustrated in Figure 2 (a). During training, the parallel decoders boost each other through sharing decoder parameters and using different object queries. On the other hand, using more groups of object queries resembles data augmentation, and behaves as query augmentation. It introduces more supervision and improve the decoder training. In addition, it is empirically observed that the encoder training is also improved, presumably with the help of the improved decoder.

Group DETR is versatile and is applicable to various DETR variants. Extensive experiments demonstrate that our approach is effective in achieving fast training convergence, shown in Figure 1. Group DETR obtains consistent improvements on various DETR-based methods [41, 36, 29, 65]. For instance, Group DETR significantly improves

Conditional DETR-C5 by 5.0 mAP with 12-epoch training on COCO [34]. The non-trivial improvements hold when we adopt longer training schedules (e.g., 36 epochs and 50 epochs). Furthermore, Group DETR outperforms baseline methods for multi-view 3D object detection [37, 38] and instance segmentation [9].

2. Background

DETR Architecture. DETR [4] is composed of an encoder, a transformer decoder, and object class and box position predictors. The encoder takes an image \mathbf{I} as input, and outputs the image feature \mathbf{X} ,

$$\text{Encoder}(\mathbf{I}) \rightarrow \mathbf{X}. \quad (1)$$

The decoder receives the image feature \mathbf{X} and the *object queries*, denoted by a matrix $\mathbf{Q} (= [\mathbf{q}_1 \mathbf{q}_2 \dots \mathbf{q}_N])$ as input, and outputs the embeddings $\tilde{\mathbf{Q}}$, followed by the predictors with the output denoted by $\mathbf{Y} (= [\mathbf{y}_1 \mathbf{y}_2 \dots \mathbf{y}_N])$,

$$\text{Decoder}(\mathbf{X}, \mathbf{Q}) \rightarrow \tilde{\mathbf{Q}}, \text{Predictor}(\tilde{\mathbf{Q}}) \rightarrow \mathbf{Y}. \quad (2)$$

The decoder is a sequence of multiple layers. Each layer includes: (i) self-attention over object queries, which performs interactions among queries for collecting the information about duplicate detection; (ii) cross-attention between queries and image features, which collects the information from image features that is useful for object detection; (iii) feed-forward network that processes the queries separately to benefit object detection.

DETR Training. The predictions during DETR training are in the set form, and have no correspondence to the ground-truth objects. DETR uses one-to-one assignment, i.e., one ground-truth object is assigned to one predictions and vice versa, through building a bipartite matching between the predictions and the ground-truth objects:

$$(\mathbf{y}_{\sigma(1)}, \bar{\mathbf{y}}_1), (\mathbf{y}_{\sigma(2)}, \bar{\mathbf{y}}_2), \dots, (\mathbf{y}_{\sigma(N)}, \bar{\mathbf{y}}_N). \quad (3)$$

Here, $\sigma(\cdot)$ is the optimal permutation of N indices, and $[\bar{\mathbf{y}}_1 \bar{\mathbf{y}}_2, \dots \bar{\mathbf{y}}_N] = \bar{\mathbf{Y}}$ corresponds to ground truth. The loss is then formulated as below:

$$\mathcal{L} = \sum_{n=1}^N \ell(\mathbf{y}_{\sigma(n)}, \bar{\mathbf{y}}_n), \quad (4)$$

where $\ell(\cdot)$ is a combination of the classification loss and the box regression loss between the ground-truth object $\bar{\mathbf{y}}$ and the prediction \mathbf{y} [4, 70, 41].

Optimization with one-to-one assignment aims to score the predictions for promoting one prediction for one ground-truth object, and demoting duplicate predictions. Such scoring needs the comparison of one prediction with other predictions, and the information of other predictions is provided from decoder self-attention over queries. The

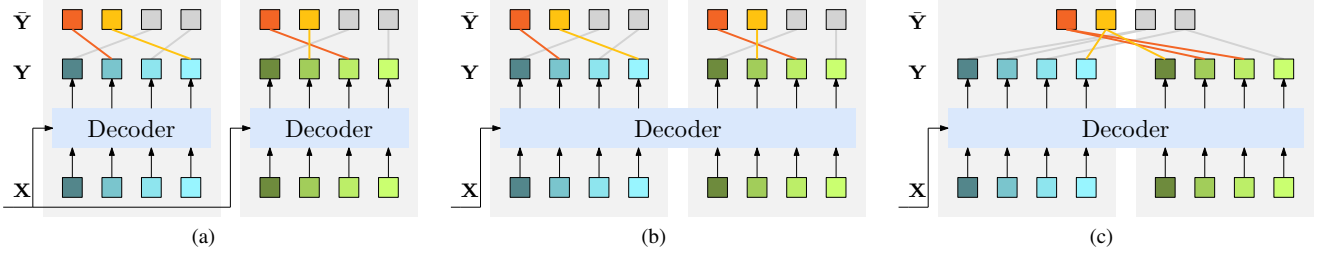


Figure 2. **Architecture illustration.** (a) Our Group DETR: group-wise one-to-many assignment and separate self-attention, architecturally equivalent to parallel decoder. (b) Group-wise one-to-many assignment only. (c) Naive one-to-many assignment. We use two groups of 4 object queries as an example. X : image features; Y : predictions; \bar{Y} : ground-truth objects, where two color boxes mean two objects and two gray boxes mean dummy objects (no objects). The color lines between Y and \bar{Y} correspond to the assignment for ground-truth objects, and the gray lines for dummy objects. For clarity, the predictors are not explicitly included.

two designs, *one-to-one assignment and self-attention over object queries*, are critical for end-to-end detection without the need of the post-processing NMS.

One-to-many assignment for non-end-to-end detection. One-to-many assignment is successfully adopted for introducing more supervision to non-end-to-end detection training, such as Faster R-CNN [44], FCOS [52], and so on [21, 33, 43, 67, 18, 6, 19]. One ground-truth object is assigned to multiple anchors or multiple pixels. During inference, a post-processing NMS is conducted for duplicate detection removal.

3. Group DETR

3.1. Algorithm

Naive one-to-many assignment. We start from a naive way for one-to-many assignment depicted in Figure 2 (c). We replace one-to-one assignment with one-to-many assignment: assign one ground-truth object to multiple predictions. It does not work and the performance is much low. The reason is that the model is trained to output multiple predictions for one ground-truth object, and lacks the scoring mechanism to promote one single prediction and demote duplicate predictions for one ground-truth object.

Group-wise one-to-many assignment. We adopt the multi-group object query mechanism: form the initial N queries as the primary group and introduce more $(K - 1)$ groups of N queries, totally K groups, $\{Q_1, Q_2, \dots, Q_K\}$. Accordingly, we have K groups of predictions, $\{Y_1, Y_2, \dots, Y_K\}$. We perform one-to-one assignment for each group, and find a bipartite matching $\sigma_k(\cdot)$, between each group of predictions and the ground-truth objects (Y_k, \bar{Y}). This results in that only one prediction for one ground-truth object is expected to score higher, and duplicate predictions is expected to score lower within one group other than within all the groups.

Separate self-attention. One-to-one assignment in one group means that the prediction assigned to one ground-

Algorithm 1 Pseudocode of one Group Decoder Layer

```
# SA: Self-Attention in the decoder layer
# CA: Cross-Attention in the decoder layer
# FFN: FFN in the decoder layer
# X: output image features of the encoder
# Q: object queries, with size (KxN, B, C)
# N, K, B, C: object query number, group number,
#           batch size, feature dimension

# group decoder
if training:
    # split object queries to K groups
    Q_list = Q.split(N, dim=0) # a list of K tensors
    parallel_Q = cat(Q_list, dim=1) # (N, KxB, C)

    # parallel self-attention
    out = SA(parallel_Q) # (N, KxB, C)
    # concat all groups: (KxN, B, C)
    out = cat(out.split(B, dim=1), dim=0)

    # cross-attention and ffn
    out = FFN(CA(out, X))
else:
    # in inference, only one group is kept
    Q = Q[:N] # (N, B, C)

    # self-attention, cross-attention, and ffn
    out = SA(Q)
    out = FFN(CA(out, X))
```

truth object is superior to other predictions within the same group. This implies that we only need to collect the information of the predictions only from the same group, rather than from all the groups. Thus we perform self-attention (abbreviated as SA) over queries for each group separately:

$$SA(Q_1), SA(Q_2), \dots, SA(Q_K). \quad (5)$$

Training architecture. The resulting architecture for training is very simple: the encoder keeps the same, and the decoder contains K separate parallel decoders as shown in Figure 2 (a):

$$\begin{aligned} \text{Decoder}(X, Q_1) &\rightarrow Q_1, \text{ Predictor}(Q_1) \rightarrow Y_1, \\ \text{Decoder}(X, Q_2) &\rightarrow Q_2, \text{ Predictor}(Q_2) \rightarrow Y_2, \\ &\dots \dots \\ \text{Decoder}(X, Q_K) &\rightarrow Q_K, \text{ Predictor}(Q_K) \rightarrow Y_K. \end{aligned} \quad (6)$$

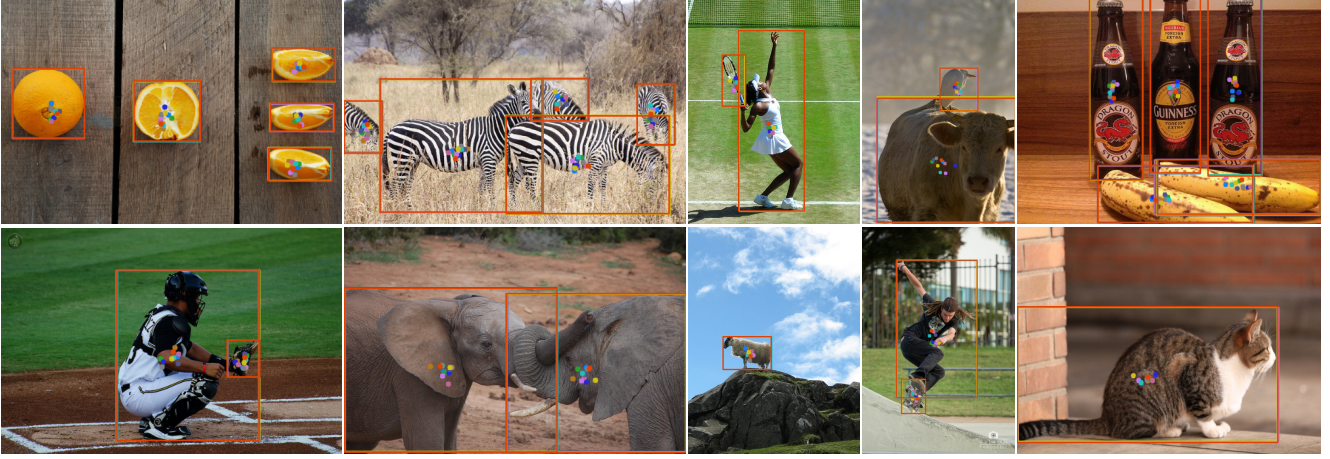


Figure 3. **Illustrating object queries.** The predicted boxes and reference points corresponding to object queries in different groups for the same ground-truth object are plotted in different colors with one color for one group. It can be seen that these queries are spatially close and can be viewed as an augmentation of other queries. The results are from Group DETR over Conditional DETR-R50 [41]. The predicted boxes and reference points may overlap. Best view in color and zoom in.

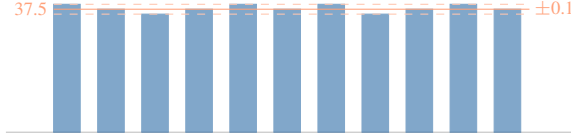


Figure 4. **The performance across groups of queries are similar.** Only a ± 0.1 mAP is observed over the median (37.5 mAP). The mAP scores over the COCO *val2017* are reported by a 12-epoch trained Conditional DETR-R50 with Group DETR.

Here, the parameters of the decoder and the predictor for the K groups are shared. Decoder separation and parallelism are feasible in that there is no interaction among queries for the other two operations, cross-attention and FFN. Our approach is called Group Decoder. In model inference, the process is the same as DETR trained normally and only needs one group of queries without any architecture modification. The pseudo-code is shown in Algorithm 1.

Loss function. The loss is an aggregation of K losses, each for one decoder. It is written as follows,

$$\mathcal{L} = \frac{1}{K} \sum_{k=1}^K \mathcal{L}_k = \frac{1}{K} \sum_{k=1}^K \sum_{n=1}^N \ell(\mathbf{y}_{\sigma_k(n)}, \bar{\mathbf{y}}_{kn}), \quad (7)$$

where $\sigma_k(\cdot)$ is the optimal permutation of N indices for the k th decoder.

3.2. Analysis

Explanation with parameter-shared models. We discuss Group DETR from the perspective of training multiple models with parameter sharing. Training with Group DETR can be regarded as simultaneously training K DETR models, which share the parameters of the encoder, the decoder, and the predictor, and only differ in the initialization

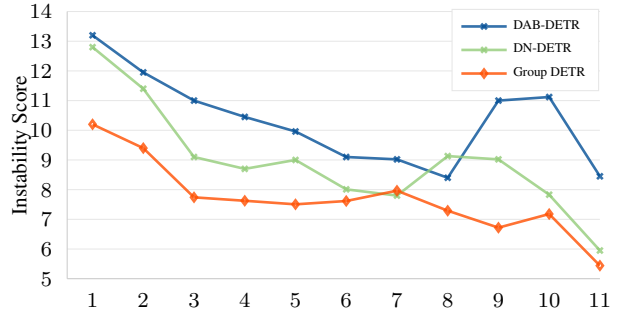


Figure 5. **More stable assignment.** The x -axis corresponds to #epoch, and the y -axis corresponds to instability score (the score is introduced by DN-DETR [29], the lower the instability score, the more stable the label assignment) over COCO *val2017*. One can see that the assignment in Group DETR is more stable than DN-DETR and its baseline DAB-DETR.

of object queries. This leads to the shared parameters receive more back-propagated gradients. Thus, these parameters are better trained and accordingly the training process converges faster.

As a side benefit, we observe that Group DETR makes the assignment more stable, as shown in Figure 5. We speculate that the stability is because the improved network leads to more reliable predictions, and thus the assignment quality is better.

Explanation with object query augmentation. The multi-group object query mechanism introduces additional $(K - 1)$ group of queries, which can be regarded as an augmentation of the primary group of queries. This is empirically illustrated in Figure 3. The reference points predicting the same objects are spatially close, and thus the corresponding object queries are similar. This may suggest that the multi-

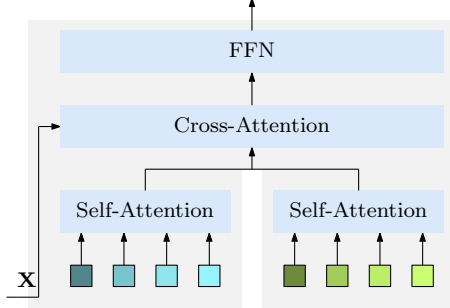


Figure 6. The parallel decoders in Group DETR are efficiently implemented as parallel self-attention, cross attention and FFN.

group object query mechanism resembles data augmentation, and at each iteration, more automatically-learned augmented queries are included, which equivalently introduces more supervision for decoder training. The results in Figure 4 empirically suggest that different groups of augmented queries lead to similar results.

The point about more supervision is also observed from the comparison between Equation 6 (for training with Group DETR) and Equation 2 (for normal DETR training). Group DETR training includes K pairs of image feature and object query group $\{(\mathbf{X}, \mathbf{Q}_1), (\mathbf{X}, \mathbf{Q}_2), \dots, (\mathbf{X}, \mathbf{Q}_K)\}$, and thus the loss contains more components as shown in Equation 7.

Table 1. **Illustrating that training with Group DETR improves both encoder and decoder.** The encoder, including CNN and transformer encoders, is initialized from a trained Conditional DETR-R50 [41] with 50 epochs and the decoder is random initialized. (a) (Fixed, Single) = the encoder is not retrained, and the decoder is trained normally without using Group DETR. (b) (Fixed, Group) = the encoder is not retrained, and the decoder is with Group DETR. (c) (Group, Group) = the encoder and the decoder are trained with Group DETR. All the results are got through training with 50 epochs. (c) > (b) implies that Group DETR also improves the encoder training.

| | Encoder | Decoder | mAP | AP_s | AP_m | AP_l |
|-----|---------|---------|-------|--------|--------|--------|
| (a) | Fixed | Single | 40.6* | 20.2 | 44.0 | 59.3 |
| (b) | Fixed | Group | 41.5 | 21.2 | 45.0 | 60.2 |
| (c) | Group | Group | 42.9* | 22.2 | 46.6 | 61.6 |

*: Training Conditional DETR with a trained encoder gives slightly lower performances than the one trained regularly, even though we train all components. New hyper-parameters may need to get better results.

Encoder training improvement. The additional supervision introduces more box regression and classification supervision from more queries assigned to each ground-truth object. The gradients with more supervision are also back-propagated from the decoder to the encoder. It is presumable that the encoder also gets benefit, verified by the empirical results in Table 1.

Computation and memory complexity. Group DETR uses more decoders during training. It is expected that

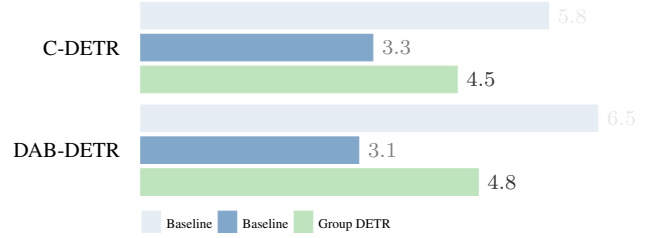


Figure 7. **Baseline models vs their Group DETR counterparts w.r.t training memory.** The gray baseline represents using the naive implementation of attention modules. With a memory-efficient implementation [12], Group DETR does not bring much memory burden during training, only requires 1.2 G and 1.7 G more GPU memory with Conditional DETR [41] (‘C-DETR’ for short) and DAB-DETR [36].

Table 2. **Group DETR outperforms baseline models with a similar training time.** Conditional DETR [41] and DAB-DETR [36] serve as baseline models to compare the performances on COCO val2017 [34]. ‘C-DETR’ and ‘w/ Group’ are the abbreviations of ‘Conditional DETR’ and ‘with Group DETR’. The entries noted by gray are the results of baseline models with the same training epochs (12 or 50 epochs) as Group DETR. To match the training times of Group DETR, we adopt longer training schedules for baselines (15 or 60 epochs). The training times are measured on 8 A100 GPUs in hours.

| Model | w/ Group | Hours | mAP | AP_s | AP_m | AP_l |
|----------|----------|-------|------|--------|--------|--------|
| C-DETR | ✓ | 4.6 | 32.6 | 14.7 | 35.0 | 48.3 |
| | | 5.8 | 34.4 | 15.1 | 37.3 | 51.3 |
| | | 5.6 | 37.6 | 18.2 | 40.7 | 55.9 |
| C-DETR | ✓ | 19.2 | 40.9 | 20.5 | 44.2 | 59.6 |
| | | 23.0 | 41.6 | 21.4 | 45.1 | 60.0 |
| | | 23.3 | 43.4 | 23.0 | 47.3 | 62.3 |
| DAB-DETR | ✓ | 5.6 | 35.2 | 16.7 | 38.6 | 51.6 |
| | | 7.0 | 36.3 | 17.1 | 39.4 | 52.5 |
| | | 6.6 | 39.1 | 19.7 | 42.5 | 56.8 |
| DAB-DETR | ✓ | 23.3 | 42.2 | 21.5 | 45.7 | 60.3 |
| | | 28.0 | 42.9 | 22.8 | 46.4 | 61.9 |
| | | 27.5 | 44.5 | 24.2 | 48.5 | 63.2 |

Group DETR will bring additional training computation costs (FLOPs) as well as training memory costs. But the parallel decoders can be implemented as a single decoder by replacing normal self-attention with parallel self-attention (depicted in Figure 6) and we can use an efficient attention implementation, FlashAttention [12, 28]. As a result, Group DETR only takes a small increase in training GPU memory and training time. For example, with Conditional DETR [41] and DAB-DETR [36], the memory increases are just 1.2 G and 1.7 G (Figure 7). The training time is increased by 5 minutes per epoch (from 23 minutes to 28 minutes and from 28 minutes to 33 minutes, respectively).

We provide the results by increasing the training time

for normal DETR training to see if Group DETR benefits simply from more training time. The results given in Table 2 show that normal training with more training time brings a little benefit and the performance is still much lower than Group DETR, implying that the performance gain from our approach is not from training time increase.

Connection to DN-DETR. DN-DETR [29] aims to stabilize one-to-one assignment during DETR training. DN-DETR forms the additional queries by adding the noises to ground-truth objects, which can be regarded as a variant of our multi-group mechanism with clear differences. In DN-DETR [29], on the one hand, the number of queries within each additional group is the same as the number of ground-truth objects. Each one correspond to one ground-truth object, and there is no query corresponding to no-object. In contrast, our approach automatically learns a number of N (e.g., 300) object queries that correspond to both ground-truth objects and no-object.

On the other hand, DN-DETR performs self-attention over noised queries, mainly for collecting the information from predictions for other objects other than from duplicate predictions. Self-attention in Group DETR instead collects both duplicate predictions and predictions for other objects.

The above two comparisons imply that DN-DETR brings the major help for the box and classification prediction, through the introduction of more positive queries corresponding to ground-truth objects (like FCOS), and no direct help for duplicate prediction removal. Our approach introduces both positive queries and negative queries (no-object), also brings the help for duplicate prediction removal.

Figure 8 shows that the performance of Group DETR is better than DN-DETR. We further investigate if Group DETR still benefits from introducing more positive queries with noised queries. As shown in Figure 8, the performance gain over Group DETR is non-trivial, a 1.5 mAP. This implies that Group DETR and DN-DETR are complementary and their major roles are different, though they have some similarities.

4. Experiments

We demonstrate the effectiveness of Group DETR in various DETR variants, and its extension to 3D detection and instance segmentation [41, 36, 29, 70, 65, 37, 38, 9]. The training setting is almost the same as baseline models, for illustrating the effectiveness of our Group DETR. We adopt the same training settings and hyper-parameters as the baseline models, such as learning rate, optimizer, pre-trained model, initialization methods, and data augmentations¹.

¹We may adjust the batch size due to the limitation of the GPU memory size for both the baseline model and our approach so that the batch size is the same.

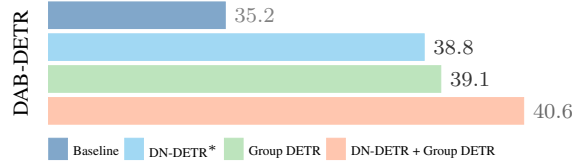


Figure 8. **Comparisons with DN-DETR.** Group DETR outperforms DN-DETR on DAB-DETR [36] (y -axis). Combining those two methods give better results, indicating they are complementary to each other. The x -axis is the mAP scores with a 12-epoch schedule on COCO *val2017*. * represents that we report the best results of DN-DETR among different numbers of denoising queries (detailed results are provided in Appendix).

4.1. Object Detection

Setting. We study various representative DETR-based detectors, such as basic baselines (Conditional DETR [41], DAB-DETR [36], DN-DETR [29]) with dense attentions, and strong baselines (DAB-Deformable-DETR [36, 70] and DINO [65, 70]) with deformable attentions. We report the results on two training schedules, training for 12 epochs and training for more epochs (36 or 50). Unless specified, the models are trained with ResNet-50 [22] as the backbone on the COCO *train2017* and evaluated on the COCO *val2017*. More implementation details are provided in Appendix.

Results. We first report the results of training with 12 epochs in Table 3. Group DETR brings consistent improvements over the baselines with dense attentions that already are superior to the original DETR [4]. It boosts Conditional DETR (-DC5) [41] by 5.0 (4.8) mAP, improves DAB-DETR (-DC5) [36] by 3.9 (4.4) mAP, and brings a 2.0 (2.6) mAP gain to DN-DETR (-DC5) [29].

Group DETR also works well on those strong baselines with deformable attentions that are equipped with two or more accelerating techniques. It gives a 1.5 mAP improvements over DAB-Deformable-DETR [36, 70]. When applying to DINO [65, 29, 70], Group DETR also exceeds it by 0.7 mAP. The gain is non-trivial over such a stronger baseline, considering that DINO is a well-tuned model² based on DAB-Deformable-DETR that combines improved hyper-parameters, improved two-stage design, improved query denoising task, and other tricks.

Furthermore, we report the results with 50 training epochs that is commonly adopted in many acceleration methods [70, 41, 8, 36]. Table 4 presents that Group DETR outperforms baseline models by large margins. For the stronger backbone, Swin-Large [40], our approach achieves 58.4 mAP (still a 0.4 mAP higher than its baseline

²In fact, our approach is compatible with query denoising and two-stage. In Table 3, for example, DN-DETR [29] utilizes query denoising, and our method improves it by 2.0. Similarly, DAB-D-DETR [36] adopts a two-stage structure, and our method achieves a 1.5 improvement.

Table 3. **Effectiveness of Group DETR with 12 epochs.** Group DETR gives consistent gains over various DETR-based baselines on COCO *val2017* [34], highlighted with brackets. All experiments adopt ResNet-50 [22] and do not use multiple patterns [57]. For DN-DETR, an improved version of DN, dynamic DN groups [65] with 100 DN queries, is used, making the results slightly different from the ones (with 3 patterns) reported in the original paper [29] (more results about the number of DN queries can be found in Appendix. ‘C-DETR’, ‘DAB-D-DETR’, and ‘w/ Group’ are ‘Conditional DETR’ [41], ‘DAB-Deformable DETR’ [36, 70], and ‘with Group DETR’, respectively, for neat representation.

| Model | w/ Group | mAP | AP _s | AP _m | AP _l |
|--------------|----------|-------------|-----------------|-----------------|-----------------|
| C-DETR | ✓ | 32.6 | 14.7 | 35.0 | 48.3 |
| | | 37.6 (+5.0) | 18.2 | 40.7 | 55.9 |
| C-DETR-DC5 | ✓ | 36.4 | 18.0 | 39.6 | 52.5 |
| | | 41.2 (+4.8) | 21.4 | 45.0 | 58.7 |
| DAB-DETR | ✓ | 35.2 | 16.7 | 38.6 | 51.6 |
| | | 39.1 (+3.9) | 19.7 | 42.5 | 56.8 |
| DAB-DETR-DC5 | ✓ | 37.5 | 19.4 | 40.6 | 53.2 |
| | | 41.9 (+4.4) | 23.3 | 45.6 | 58.4 |
| DN-DETR | ✓ | 38.6 | 17.9 | 41.6 | 57.7 |
| | | 40.6 (+2.0) | 19.8 | 43.9 | 59.4 |
| DN-DETR-DC5 | ✓ | 41.9 | 22.2 | 45.1 | 59.8 |
| | | 44.5 (+2.6) | 25.9 | 48.2 | 62.2 |
| DAB-D-DETR | ✓ | 44.2 | 27.5 | 47.1 | 58.6 |
| | | 45.7 (+1.5) | 28.1 | 49.0 | 60.6 |
| DINO-4scale | ✓ | 49.4 | 32.3 | 52.5 | 63.2 |
| | | 50.1 (+0.7) | 32.4 | 53.2 | 64.7 |

DINO [65] (58.0 mAP with Swin-Large)). This verifies the generalization ability of our Group DETR.

Last, we compare the training convergence curves of the baseline models and their Group DETR counterparts. The results, as shown in Figure 1, provide more evidence that Group DETR speeds DETR training convergence on various DETR variants.

System-level Results on COCO *test-dev* with ViT-Huge. We also have the system-level performance on COCO *test-dev* [34] with ViT-Huge [14]. We apply Group DETR to DINO [65] and follow its training pipeline and settings: pre-train the encoder with a self-supervised method, then pre-train the whole model on Object365 [46], and last fine-tune the whole model on COCO [34]. Our model is the first to achieve 64.5 mAP on COCO *test-dev*, which is still superior to other methods with larger encoder and more pre-training data [39, 55, 58, 60]. The details and comparisons with other methods are provided in Appendix.

Table 4. **Effectiveness of Group DETR with more epochs.** Group DETR still outperforms baselines by non-trivial margins with more training epochs (36 or 50 epochs). Settings and notations are consistent with Table 3, except for the training epochs (36 epochs for DINO-4scale by following the original paper [65] and 50 epochs for other models). ‘DINO-4scale-Swin-L’ means it adopts Swin-Large [40] as the backbone.

| Model | w/ Group | mAP | AP _s | AP _m | AP _l |
|--------------------|----------|-------------|-----------------|-----------------|-----------------|
| C-DETR | ✓ | 40.9 | 20.5 | 44.2 | 59.6 |
| | | 43.4 (+2.5) | 23.0 | 47.3 | 62.3 |
| C-DETR-DC5 | ✓ | 43.7 | 23.9 | 47.6 | 60.1 |
| | | 45.8 (+2.1) | 26.8 | 49.7 | 63.1 |
| DAB-DETR | ✓ | 42.2 | 21.5 | 45.7 | 60.3 |
| | | 44.5 (+2.3) | 24.2 | 48.5 | 63.2 |
| DAB-DETR-DC5 | ✓ | 44.5 | 25.3 | 48.2 | 62.3 |
| | | 46.7 (+2.2) | 27.6 | 50.9 | 64.0 |
| DN-DETR | ✓ | 44.0 | 23.9 | 47.7 | 62.9 |
| | | 45.4 (+1.4) | 25.1 | 49.3 | 63.8 |
| DN-DETR-DC5 | ✓ | 47.5 | 27.9 | 50.7 | 65.9 |
| | | 48.0 (+0.5) | 29.3 | 52.1 | 65.4 |
| DAB-D-DETR | ✓ | 48.1 | 31.4 | 51.4 | 63.4 |
| | | 49.7 (+1.6) | 31.4 | 52.5 | 65.6 |
| DINO-4scale | ✓ | 50.9 | 34.6 | 54.1 | 64.6 |
| | | 51.3 (+0.4) | 34.7 | 54.5 | 65.3 |
| DINO-4scale-Swin-L | ✓ | 58.0 | 41.3 | 61.9 | 74.0 |
| | | 58.4 (+0.4) | 41.0 | 62.5 | 73.9 |

4.2. More Applications

Group DETR is applicable to DETR-style techniques to other vision problems. We report the results for two additional problems: multi-view 3D object detection [24, 32, 37, 38] and instance segmentation [9, 30], to further demonstrate the effectiveness.

Multi-view 3D object detection. We report the results over PETR [37] and PETR v2 [38] on the nuScenes *val* dataset [2]. Table 5 shows that Group DETR brings significant gains to PETR and PETR v2 with 24 training epochs in terms of both the nuScenes Detection Score (NDS) and mAP scores.

Table 5. **Results on multi-view 3D object detection.** All experiments are evaluated on the nuScenes *val* set [2]. We train these experiments for 24 epochs with VoVNetV2 [27] as the backbone and with the image size of 800 × 320. We follow all the settings and hyper-parameters of PETR [37] and PETR v2 [38].

| Model | w/ Group | NDS | mAP |
|---------|----------|-------------|-------------|
| PETR | ✓ | 42.0 | 37.4 |
| | | 45.0 (+3.0) | 38.8 (+1.4) |
| PETR v2 | ✓ | 50.3 | 40.7 |
| | | 51.3 (+1.0) | 41.9 (+1.2) |

Instance segmentation. We demonstrate the effectiveness of the representative method, Mask2Former [9]. The results

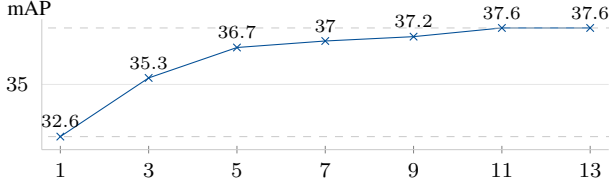


Figure 9. **Influence of group number.** The x -axis is the number of groups. It can be seen that the performance becomes stable when the number of groups reaches 11.

are given in Table 6. Group DETR achieves a 1.2 (0.3) mAP^m gain with 12 (50) epochs.

Table 6. **Results on instance segmentation.** The mask mAP (mAP^m) is used for instance segmentation on COCO *val2017*. We adopt Mask2Former [9] as the baseline. The experiments are conducted with ResNet-50 [22] as the backbone, following all the settings of Mask2Former.

| Epochs | w/ Group | mAP^m | AP_s^m | AP_m^m | AP_l^m |
|--------|----------|----------------|-----------------|-----------------|-----------------|
| 12 | | 38.5 | 17.6 | 41.4 | 60.4 |
| 12 | ✓ | 39.7 (+1.2) | 18.7 | 42.8 | 60.8 |
| 50 | | 43.7 | 23.4 | 47.2 | 64.8 |
| 50 | ✓ | 44.0 (+0.3) | 23.8 | 47.1 | 65.1 |

4.3. Ablation Study

We conduct the ablation study by using Conditional DETR [41] as the baseline. The CNN backbone is ResNet-50 [22], and the training epoch number is 12. The performances are evaluated on COCO *val2017* [34]. We mainly study the effects of the key design: group-wise one-to-many assignment, separate self-attention, and group number.

Group-wise one-to-many assignment and separate self-attention. Table 7 shows how group-wise one-to-many (o2m) assignment and separate self-attention make contributions. In comparison to the baseline (a), group-wise o2m assignment improves the mAP score from 32.6 mAP to 34.8 mAP : with the gain 2.2. The separate self attention (Sep. SA) further gets a 2.8 mAP gain. In addition, we report naive one-to-many assignment. The results are very poor, which is reasonable in that there are duplicate predictions and there is a lack of scoring mechanisms for demoting them. The results suggest that both group-wise o2m assignment and separate self-attention are effective.

Group number. Figure 9 shows the influence of the number of groups K in Group DETR. The detection performance improves when increasing the number of groups, and becomes stable when the group number reaches 11. Thus, we adopt $K = 11$ by default in Group DETR in our experiments.

Table 7. **Effects of group-wise one-to-many assignment and separate self-attention.** (a) baseline: one-to-one assignment with 300 object queries. (b) naive one-to-many assignment with 3300 object queries for training and inference. (c) group-wise one-to-many assignment and no separate self-attention with 11 groups of 300 queries, inference with a group of 300 queries. (d) group-wise one-to-many assignment and separate self-attention with 11 groups of 300 queries, inference with a group of 300 queries. o2m = one-to-many, Sep. SA = separate self-attention.

| | o2m | Sep. SA | mAP | AP_s | AP_m | AP_l |
|-----|-------|---------|--------------|---------------|---------------|---------------|
| (a) | × | × | 32.6 | 14.4 | 34.9 | 48.6 |
| (b) | Naive | × | 8.4 | 8.0 | 13.2 | 13.3 |
| (c) | Group | × | 34.8 | 16.4 | 37.7 | 51.4 |
| (d) | Group | ✓ | 37.6 | 18.2 | 40.7 | 55.9 |

5. Related Works

There are two main lines for accelerating DETR training: modify *cross-attention* and stabilize *one-to-one assignment*. The two are complementary and can be combined to further boost the performance.

Modifying cross-attention. Cross-attention module aims to collect the information from the image features useful to classification and localization. Various methods are proposed to select the informative image regions more efficiently and effectively [16, 8, 57, 61, 36, 17]. For example, Deformable attention [70] selects the highly informative positions dynamically according to the previous decoder embedding. Conditional DETR [8] instead continues to use the normal global attention, and dynamically computes the spatial attention to softly select the informative regions. SMCA [16] uses the Gaussian-like weight for spatial modulation.

Stabilizing one-to-one assignment. DETR [4] relies on one-to-one assignment, where each ground-truth object is assigned to a single prediction through building a bipartite matching between the predictions and the ground-truth objects. DN-DETR [29] finds the assignment process is unstable and attributes the slow convergence issue to the instabilities. Thus, DN-DETR [29] introduces groups of noisy queries by adding noises to ground-truth objects, to stabilize the assignment, leading to faster convergence. DINO [65] makes further improvement through contrastive denoising training to generate both positive and negative noise queries with different noise levels. Our approach studies the assignment mechanism instead for introducing more supervision.

One-to-many assignment. One-to-many assignment is widely adopted in deep detectors [44, 21, 33, 52], and has attracted a lot of interest [67, 26, 69, 18, 6, 54, 49]. For example, Faster R-CNN [44] and FCOS [52] produce multiple positive anchors and pixels for each ground-truth object. In this paper, we investigate one-to-many assignment in a fea-

sible manner for the end-to-end detector DETR.

Concurrent with our work, \mathcal{H} -DETR [25] also uses one-to-many assignment to speed up DETR training convergence. Our Group DETR and \mathcal{H} -DETR are related, but different: (1) Group DETR introduces group-wise one-to-many assignment with separate self-attention with the same number of object queries in each group. \mathcal{H} -DETR adopts hybrid assignments in two different groups: One group uses one-to-one assignment and another uses one-to-many assignment with more object queries. (2) All the decoders in Group DETR can be used for inference. But the additional decoder in \mathcal{H} -DETR is not directly used and requires NMS for inference. (3) During training, our architecture introduces one parameter: the number of groups. In contrast, \mathcal{H} -DETR introduces the number of additional queries and the number of additional positive queries.

DETA [42] is another concurrent work with our Group DETR. DETA directly uses one-to-many assignment and brings NMS back to DETR frameworks. While our method provides group-wise one-to-many assignment and maintains end-to-end detection.

6. Conclusion

The key points in Group DETR include group-wise one-to-many assignment and parallel self-attention. The success stems from involving more groups of object queries as an addition to the primary group of object queries, and thus introducing more supervision. Group-wise assignment mechanism makes sure that the competition among predictions happens within each group separately, and separate self-attention eases the training. Thus, the NMS pose-processing is not necessary, and the inference process is kept the same as normally trained DETR and not dependent on the group design. Our approach is simple, easily implemented, and general.

Acknowledgements. This work is supported by the Sichuan Science and Technology Program (2023YFSY0008), National Natural Science Foundation of China (61632003, 61375022, 61403005), Grant SCITLAB-20017 of Intelligent Terminal Key Laboratory of SiChuan Province, Beijing Advanced Innovation Center for Intelligent Robots and Systems (2018IRS11), and PEK-SenseTime Joint Laboratory of Machine Vision.

Appendix

A. More Details and Results

A.1. Datasets and Evaluation Metrics

We perform the object detection and instance segmentation experiments on the COCO 2017 [34] dataset, which contains about 118K training (*train2017*) images, 5K validation (*val2017*) images, and 20K testing (*test-dev*) im-

ages. Following the common practice, we train our model on COCO *train2017* and report the standard mean average precision (mAP) result (box mAP for object detection and mask mAP for instance segmentation) on the COCO *val2017* dataset under different IoU thresholds (from 0.5 to 0.95) and object scales (small, medium, and large). We also report the result on COCO *test-dev* with a large foundation model (ViT-Huge [62, 20, 7]).

We perform multi-view 3D object detection experiments on the nuScenes [2] dataset, which contains 1000 driving sequences. There are 700 for *train* set, 150 for *val* set and 150 for *test* set. We report the standard nuScenes Detection Score (NDS) and mean Average Precision (mAP) result on the nuScenes *val* set.

A.2. Implementation Details

Our Group DETR adopts multiple groups of object queries. Each group shares the same architectures and numbers of object queries³. It resembles data augmentation with automatically-learned object query augmentation and is also equivalent to simultaneously training parameter-sharing networks of the same architecture.

In one-stage DETR frameworks, including Conditional DETR [41], DAB-DETR [36], DN-DETR [29], and DAB-Deformable-DETR [36, 70], we can easily implement Group DETR by adopting multiple groups of learnable object queries. While the situation is different in two-stage DETR frameworks, such as DINO [65]. The initializations of object queries are dependent on the top- N predicted boxes of the first stage. To make the object queries in multiple groups similar to each other, we construct multiple pairs of classification and regression prediction heads in the first stage, each pair of which provides initialization for the object queries in the corresponding group. As for model inference, we only need one pair of these prediction heads, the same as the original model.

A.3. More Results of DN-DETR

Results of DN-DETR with different numbers of denoising queries. We conduct experiments with different numbers of denoising queries in DN-DETR [29]. The results in Figure 10 suggest that increasing the number of denoising queries can not achieve further improvements and show unstable performances. The effects of denoising queries differ from the ones of Group DETR (Figure 8 in the main paper). We choose to use 100 denoising queries in our experiments in Table 3 and Table 4 in the main paper by following the setting in the original paper [29]. To make direct comparisons with DN-DETR [29], we report the best results across

³When applying Group DETR to DN-DETR [29] and DINO [65], we add the corresponding query denoising task in each group to keep the same architecture with the original implementation.

Table 8. Our method achieves 64.5 mAP on the COCO test-dev.

| Method | #Params | Encoder Pretraining Data | Detector Pretraining Data | w/ Mask | mAP |
|--|---------|--------------------------------|---------------------------|---------|------|
| Swin-L (HTC++) [40] | 284M | IN-22K (14M) | n/a | ✓ | 58.7 |
| DyHead (Swin-L) [11] | 213M | IN-22K (14M) | n/a | ✓ | 60.6 |
| Soft-Teacher (Swin-L) [59] | 284M | IN-22K (14M) | COCO-unlabeled + O365 | ✓ | 61.3 |
| GLIP (DyHead) [31] | ≥284M | IN-22K (14M) | FourODs + GoldG + Cap24M | × | 61.5 |
| Florence (CoSwin-H) [66] | ≥637M | FLD-900M (900M) | FLD-9M | × | 62.4 |
| GLIPv2 (CoSwin-H) [66] | ≥637M | FLD-900M (900M) | merged data ^b | ✓ | 62.4 |
| SwinV2-G (HTC++) [39] | 3.0B | IN-22K + ext-70M (84M) | O365 | ✓ | 63.1 |
| DINO-5scale (Swin-L) [65] | 218M | IN-22K (14M) | O365 | × | 63.3 |
| BEIT-3 (ViTDet) [55] | 1.9B | merged data ^a | O365 | ✓ | 63.7 |
| FD-SwinV2-G (HTC++) [58] | 3.0B | IN-22K + IN-1K + ext-70M (85M) | O365 | ✓ | 64.2 |
| FocalNet-H (DINO-5scale) [60] | 746M | IN-22K (14M) | O365 | × | 64.3 |
| Co-Deformable-DETR (MixMIM-g) [35, 71] | 1.0B | IN-1K (1M) | O365 | × | 64.5 |
| EVA (CMask R-CNN) [15, 3, 21] | ≥1.0B | merged-30M ^c | O365 | ✓ | 64.7 |
| InternImage-H (DINO-5scale) [56, 48, 65] | 2.18B | merged data ^d | O365 | × | 65.4 |
| ViT-Huge + Group DETR (DINO-4scale) | 629M | IN-1K (1M) | O365 | × | 64.5 |

All the results are achieved with test time augmentation. In the table, we follow the notations for various datasets used in DINO [65] and FocalNet [60].

‘w/ Mask’ means using mask annotations when finetuning the detectors on COCO [34]. And for the baseline DINO, we adopt the 4scale version [65].

‘merged data^a’: IN-22K + Image-Text (35M) + Text (160GB). ‘merged data^b’: FourODs + INBoxes + GoldG + CC15M + SBU.

‘merged-30M^c’: IN-21K + O365 + COCO + ADE20K + CC15M. ‘merged data^d’: Laion-400M + YFCC-15M + CC12M.

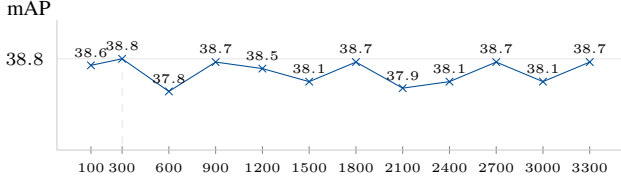


Figure 10. Results of DN-DETR with different number of denoising queries. We show the detection performances (mAP) on MS COCO [34] of adopting different number of denoising queries in DN-DETR.

different numbers of denoising queries in Figure 10 (38.8 mAP).

A.4. Applying Group DETR to SAM-DETR series

We also apply Group DETR to another stream of work to accelerate DETR training, SAM-DETR [63] and SAM-DETR++ [64]. The results are given in Table 9. Improvements on SAM-DETR [63] (gains: 3.1 mAP with 12e and 1.9 mAP with 50e) and SAM-DETR++ [64] (gains: 2.2 mAP with 12e and 1.3 mAP with 50e) show that Group DETR is complementary to them as well.

B. More Comparisons on COCO test-dev

Settings. To compare state-of-the-art results on COCO test-dev, we follow DINO [65] to build our model with a large foundation model, ViT-Huge. We follow its training pipeline and settings: (i) pre-train [7] and fine-tune the ViT-Huge on ImageNet-1K [13], (ii) pre-train the whole detector on Object365 [46] for 24 epochs with 64 A100 GPUs,

and (iii) finetune the detector on COCO [34] for 20 epochs with 32 A100 GPUs. When pre-training the detector on Object365, we follow DINO [65] to only leave the first 5k out of 80k validation images as the validation set and add the other images to the training set. We also use other schemes when training the detector on Object365 and COCO, such as enlarging the image size to 1.5× when finetuning and adopting test time augmentation. In addition, we apply the exponential moving average (EMA) technique [50], use CDN queries [65], and adopt 11 groups with Group DETR during detector pre-training and fine-tuning. When finetuning the detector on COCO, we find that applying learning rate decay [10, 1, 20, 7] for the components of the detector gives a ~0.9 mAP gain on COCO. During testing, we adopt test time augmentation with various scales and their flipped counterparts and perform fusion⁴ on the query features and the final predictions [65].

Results. Table 8 shows the results. Our model is the first to achieve 64.5 mAP on COCO test-dev. Only pre-training the ViT-Huge on ImageNet-1K [13], our model can outperform other methods with larger models (e.g., BEIT-3 [55] and SwinV2-G [39, 58]) and more pre-training data. Models such as EVA [15] and InterImage-H [56], with larger foundation models (ViT-giant [62] or InterImage-H [56]) and more data [13, 5, 47, 68, 51, 45], give higher results (64.7 mAP and 65.4 mAP) than our model. We expect that our results will be further improved with more pre-training data and larger models.

⁴According to our experiments, the fusion on the query features builds a robust feature across different scales and gives a ~0.8 mAP improvement.

Table 9. **Effectiveness of Group DETR on SAM-DETR and SAM-DETR++**. All experiments adopt ResNet-50 [22] and evaluate on COCO val2017 [34].

| Model | w/ Group | Epochs | mAP |
|------------|----------|--------|-------------|
| SAM-DETR | ✓ | 12 | 33.1 |
| | | 12 | 36.2 (+3.1) |
| SAM-DETR | ✓ | 50 | 39.8 |
| | | 50 | 41.7 (+1.9) |
| SAM-DETR++ | ✓ | 12 | 41.1 |
| | | 12 | 43.3 (+2.2) |
| SAM-DETR++ | ✓ | 50 | 46.1 |
| | | 50 | 47.4 (+1.3) |

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 10
- [2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 7, 9
- [3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019. 10
- [4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 1, 2, 6, 8
- [5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568, 2021. 10
- [6] Qiang Chen, Yingming Wang, Tong Yang, Xiangyu Zhang, Jian Cheng, and Jian Sun. You only look one-level feature. In *CVPR*, pages 13039–13048, 2021. 3, 8
- [7] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *CoRR*, abs/2202.03026, 2022. 9, 10
- [8] Xiaokang Chen, Fangyun Wei, Gang Zeng, and Jingdong Wang. Conditional detr v2: Efficient detection transformer with box queries. *arXiv preprint arXiv:2207.08914*, 2022. 1, 6, 8
- [9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *arXiv preprint arXiv:2112.01527*, 2021. 2, 6, 7, 8
- [10] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*, 2020. 10
- [11] Xiyang Dai, Yinpeng Chen, Jianwei Yang, Pengchuan Zhang, Lu Yuan, and Lei Zhang. Dynamic detr: End-to-end object detection with dynamic attention. In *ICCV*, pages 2988–2997, 2021. 10
- [12] Tri Dao, Daniel Y Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *arXiv preprint arXiv:2205.14135*, 2022. 5
- [13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 10
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 7
- [15] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 10
- [16] Peng Gao, Minghang Zheng, Xiaogang Wang, Jifeng Dai, and Hongsheng Li. Fast convergence of detr with spatially modulated co-attention. In *ICCV*, pages 3621–3630, 2021. 1, 8
- [17] Ziteng Gao, Limin Wang, Bing Han, and Sheng Guo. Adamixer: A fast-converging query-based object detector. In *CVPR*, pages 5364–5373, 2022. 1, 8
- [18] Zheng Ge, Songtao Liu, Zeming Li, Osamu Yoshie, and Jian Sun. Ota: Optimal transport assignment for object detection. In *CVPR*, pages 303–312, 2021. 3, 8
- [19] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 3
- [20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, June 2022. 9, 10
- [21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 3, 8, 10
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1, 6, 7, 8, 11
- [23] J Hosang, R Benenson, and B Schiele. Learning non-maximum suppression. *PAMI*, 2017. 1
- [24] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 7
- [25] Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, and Han Hu. Dets with hybrid matching. *arXiv preprint arXiv:2207.13080*, 2022. 9
- [26] Kang Kim and Hee Seok Lee. Probabilistic anchor assignment with iou prediction for object detection. In *ECCV*, pages 355–371. Springer, 2020. 8

- [27] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, pages 13906–13915, 2020. 7
- [28] Benjamin Lefaudeaux, Francisco Massa, Diana Liskovich, Wenhan Xiong, Vittorio Caggiano, Sean Naren, Min Xu, Jieru Hu, Marta Tintore, Susan Zhang, Patrick Labatut, and Daniel Haziza. xformers: A modular and hackable transformer modelling library. <https://github.com/facebookresearch/xformers>, 2022. 5
- [29] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *CVPR*, 2022. 2, 4, 6, 7, 8, 9
- [30] Feng Li, Hao Zhang, Shilong Liu, Lei Zhang, Lionel M Ni, Heung-Yeung Shum, et al. Mask dino: Towards a unified transformer-based framework for object detection and segmentation. *arXiv preprint arXiv:2206.02777*, 2022. 7
- [31] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 10
- [32] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 7
- [33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 1, 3, 8
- [34] 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 *ECCV*, 2014. 1, 2, 5, 7, 8, 9, 10, 11
- [35] Jihao Liu, Xin Huang, Yu Liu, and Hongsheng Li. Mixmim: Mixed and masked image modeling for efficient visual representation learning. *arXiv preprint arXiv:2205.13137*, 2022. 10
- [36] 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, 2, 5, 6, 7, 8, 9
- [37] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022. 2, 6, 7
- [38] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022. 2, 6, 7
- [39] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019, 2022. 7, 10
- [40] 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. *arXiv:2103.14030*, 2021. 6, 7, 10
- [41] Depu Meng, Xiaokang Chen, ZeJia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *ICCV*, pages 3651–3660, 2021. 1, 2, 4, 5, 6, 7, 8, 9
- [42] Jeffrey Ouyang-Zhang, Jang Hyun Cho, Xingyi Zhou, and Philipp Krähenbühl. Nms strikes back. *arXiv preprint arXiv:2212.06137*, 2022. 9
- [43] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1, 3
- [44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 1, 2, 3, 8
- [45] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 10
- [46] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8430–8439, 2019. 7, 10
- [47] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 10
- [48] Weijie Su, Xizhou Zhu, Chenxin Tao, Lewei Lu, Bin Li, Gao Huang, Yu Qiao, Xiaogang Wang, Jie Zhou, and Jifeng Dai. Towards all-in-one pre-training via maximizing multi-modal mutual information. *arXiv preprint arXiv:2211.09807*, 2022. 10
- [49] Peize Sun, Yi Jiang, Enze Xie, Wenqi Shao, Zehuan Yuan, Changhu Wang, and Ping Luo. What makes for end-to-end object detection? In *ICML*, pages 9934–9944. PMLR, 2021. 8
- [50] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 10
- [51] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016. 10
- [52] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *ICCV*, pages 9627–9636, 2019. 2, 3, 8
- [53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1
- [54] Jianfeng Wang, Lin Song, Zeming Li, Hongbin Sun, Jian Sun, and Nanning Zheng. End-to-end object detection with

- fully convolutional network. In *CVPR*, pages 15849–15858, 2021. 8
- [55] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 7, 10
- [56] Wenhui Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, et al. Internimage: Exploring large-scale vision foundation models with deformable convolutions. *arXiv preprint arXiv:2211.05778*, 2022. 10
- [57] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *AAAI*, 2022. 1, 7, 8
- [58] Yixuan Wei, Han Hu, Zhenda Xie, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv preprint arXiv:2205.14141*, 2022. 7, 10
- [59] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3060–3069, 2021. 10
- [60] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks, 2022. 7, 10
- [61] Zhuyu Yao, Jiangbo Ai, Boxun Li, and Chi Zhang. Efficient detr: improving end-to-end object detector with dense prior. *arXiv preprint arXiv:2104.01318*, 2021. 1, 8
- [62] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12104–12113, 2022. 9, 10
- [63] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating detr convergence via semantic-aligned matching. In *CVPR*, pages 949–958, 2022. 10
- [64] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Jiaxing Huang, Kaiwen Cui, Shijian Lu, and Eric P Xing. Semantic-aligned matching for enhanced detr convergence and multi-scale feature fusion. *arXiv preprint arXiv:2207.14172*, 2022. 10
- [65] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2, 6, 7, 8, 9, 10
- [66] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. *arXiv preprint arXiv:2206.05836*, 2022. 10
- [67] 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 *CVPR*, pages 9759–9768, 2020. 3, 8
- [68] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 10
- [69] Benjin Zhu, Jianfeng Wang, Zhengkai Jiang, Fuhang Zong, Songtao Liu, Zeming Li, and Jian Sun. Autoassign: Differentiable label assignment for dense object detection. *arXiv preprint arXiv:2007.03496*, 2020. 8
- [70] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. *CoRR*, abs/2010.04159, 2020. 1, 2, 6, 7, 8, 9
- [71] Zhuofan Zong, Guanglu Song, and Yu Liu. Detsr with collaborative hybrid assignments training. *arXiv preprint arXiv:2211.12860*, 2022. 10