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

Qiang Chen[1*], Xiaokang Chen[2*], Jian Wang[1], Shan Zhang[3]
Kun Yao[1], Haocheng Feng[1], Junyu Han[1], Errui Ding[1], Gang Zeng[2], Jingdong Wang[1†]
[1]Baidu VIS
[2]Key Lab. of Machine Perception (MoE), School of IST, Peking University
[3]Australian National University
{chenqiang13,wangjian33}@baidu.com
{fenghaocheng,hanjunyu,dingerrui,wangjingdong}@baidu.com
{pkucxk,gang.zeng}@pku.edu.cn, shan.zhang@anu.edu.au

## A. More Details and Results

### A.1. Datasets and Evaluation Metrics

We perform the object detection and instance segmentation experiments on the COCO 2017 [15] dataset, which contains about 118K training (*train2017*) images, 5K validation (*val2017*) images, and 20K testing (*test-dev*) images. 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 [32, 10, 5]).

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 queries[1]. 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 [20], DAB-DETR [17], DN-DETR [13], and DAB-Deformable-DETR [17, 38], 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 [35]. 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 [13]. 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 [13]. To make direct comparisons with DN-DETR [13], we report the best results across 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 [33] and SAM-DETR++ [34]. The results are given in Table 9. Improvements on SAM-DETR [33] (gains: 3.1 mAP with 12e and 1.9 mAP with 50e) and SAM-DETR++ [34] (gains: 2.2 mAP with 12e and 1.3 mAP with 50e) show that Group

---

*Equal contribution.
†Corresponding author.
[1]When applying Group DETR to DN-DETR [13] and DINO [35], 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++) [19] | 284M | IN-22K (14M) | n/a | ✓ | 58.7 |
| DyHead (Swin-L) [7] | 213M | IN-22K (14M) | n/a | ✓ | 60.6 |
| Soft-Teacher (Swin-L) [30] | 284M | IN-22K (14M) | COCO-unlabeled + O365 | ✓ | 61.3 |
| GLIP (DyHead) [14] | ≥284M | IN-22K (14M) | FourODs + GoldG + Cap24M | × | 61.5 |
| Florence (CoSwin-H) [36] | ≥637M | FLD-900M (900M) | FLD-9M | × | 62.4 |
| GLIPv2 (CoSwin-H) [36] | ≥637M | FLD-900M (900M) | merged data$^b$ | ✓ | 62.4 |
| SwinV2-G (HTC++) [18] | 3.0B | IN-22K + ext-70M (84M) | O365 | ✓ | 63.1 |
| DINO-5*scale* (Swin-L) [35] | 218M | IN-22K (14M) | O365 | × | 63.3 |
| BEIT-3 (ViTDet) [27] | 1.9B | merged data$^a$ | O365 | ✓ | 63.7 |
| FD-SwinV2-G (HTC++) [29] | 3.0B | IN-22K + IN-1K + ext-70M (85M) | O365 | ✓ | 64.2 |
| FocalNet-H (DINO-5*scale*) [31] | 746M | IN-22K (14M) | O365 | × | 64.3 |
| *Co*-Deformable-DETR (MixMIM-g) [16, 39] | 1.0B | IN-1K (1M) | O365 | × | 64.5 |
| EVA (CMask R-CNN) [9, 3, 11] | ≥1.0B | merged-30M$^c$ | O365 | ✓ | 64.7 |
| InternImage-H (DINO-5*scale*) [28, 24, 35] | 2.18B | merged data$^d$ | O365 | × | 65.4 |
| **ViT-Huge + Group DETR (DINO-4*scale*)** | 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 [35] and FocalNet [31]. 'w/ Mask' means using mask annotations when finetuning the detectors on COCO [15]. And for the baseline DINO, we adopt the 4*scale* version [35]. '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.

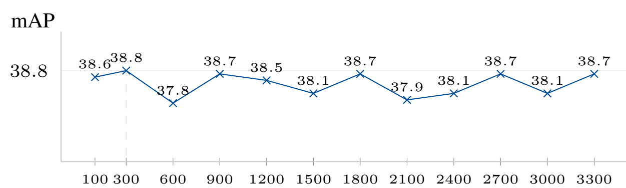


Figure 10. **Results of DN-DETR with different number of denoising queries.** We show the detection performances (mAP) on MS COCO [15] of adopting different number of denoising queries in DN-DETR.

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 [35] to build our model with a large foundation model, ViT-Huge. We follow its training pipeline and settings: (i) pre-train [5] and fine-tune the ViT-Huge on ImageNet-1K [8], (ii) pre-train the whole detector on Object365 [22] for 24 epochs with 64 A100 GPUs, and (iii) finetune the detector on COCO [15] for 20 epochs with 32 A100 GPUs. When pre-training the detector on Object365, we follow DINO [35] 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 [25], use CDN queries [35], 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 [6, 1, 10, 5] 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 fusion[2] on the query features and the final predictions [35].

Table 9. **Effectiveness of Group DETR on SAM-DETR and SAM-DETR++.** All experiments adopt ResNet-50 [12] and evaluate on COCO *val2017* [15].

| 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) |

**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 [8], our model can outperform other methods with larger models (e.g., BEIT-3 [27] and SwinV2-G [18, 29]) and more pre-training data. Models such as EVA [9] and InterImage-H [28], with larger

[2] According to our experiments, the fusion on the query features builds a robust feature across different scales and gives a ∼0.8 mAP improvement.

foundation models (ViT-giant [32] or InterImage-H [28]) and more data [8, 4, 23, 37, 26, 21], 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.

## References


[1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[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. 1

[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. 2

[4] 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. 3

[5] 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. 1, 2

[6] 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. 2

[7] 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. 2

[8] 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. 2, 3

[9] 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. 2

[10] 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. 1, 2

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[13] 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. 1

[14] 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. 2

[15] 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

[16] 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. 2

[17] 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

[18] 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. 2

[19] 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. 2

[20] 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

[21] 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. 3

[22] 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. 2

[23] 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. 3

[24] 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. 2

[25] 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. 2

[26] 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. 3

[27] 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. 2

[28] Wenhai 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. 2, 3

[29] 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. 2

[30] 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. 2

[31] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks, 2022. 2

[32] 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. 1, 3

[33] Gongjie Zhang, Zhipeng Luo, Yingchen Yu, Kaiwen Cui, and Shijian Lu. Accelerating detr convergence via semantic-aligned matching. In *CVPR*, pages 949–958, 2022. 1

[34] 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. 1

[35] 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. 1, 2

[36] 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. 2

[37] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *CVPR*, 2017. 3

[38] 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

[39] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. *arXiv preprint arXiv:2211.12860*, 2022. 2