Group DETR v2: Strong Object Detector with Encoder-Decoder Pretraining

Qiang Chen[1*], Jian Wang[1*], Chuchu Han[1*], Shan Zhang[2], Zexian Li[3], Xiaokang Chen[4], Jiahui Chen[3]
Xiaodi Wang[1], Shuming Han[1], Gang Zhang[1], Haocheng Feng[1], Kun Yao[1], Junyu Han[1], Errui Ding[1]
Jingdong Wang[1†]

[1]Baidu VIS [2]Australian National University [3]Beihang University [4]Peking University

Table 1. **Our method establishes a new SoTA on the COCO *test-dev* leaderboard.**

Method	#Params	Encoder Pretraining Data	Detector Pretraining Data	w/ Mask	mAP
Swin-L (HTC++) [16]	284M	IN-22K (14M)	n/a	✓	58.7
DyHead (Swin-L) [6]	213M	IN-22K (14M)	n/a	✓	60.6
Soft-Teacher (Swin-L) [25]	284M	IN-22K (14M)	COCO-unlabeled + O365	✓	61.3
GLIP (DyHead) [11]	$\geq$284M	IN-22K (14M)	FourODs + GoldG + Cap24M	×	61.5
Florence (CoSwin-H) [29]	$\geq$637M	FLD-900M (900M)	FLD-9M	×	62.4
GLIPv2 (CoSwin-H) [29]	$\geq$637M	FLD-900M (900M)	FourODs + INBoxes + GoldG + CC15M + SBU	✓	62.4
SwinV2-G (HTC++) [15]	3.0B	IN-22K + ext-70M (84M)	O365	✓	63.1
DINO (Swin-L) [28]	218M	IN-22K (14M)	O365	×	63.3
BEIT-3 (ViTDet) [22]	1.9B	IN-22K + Image-Text (35M) + Text (160GB)	O365	✓	63.7
FD-SwinV2-G (HTC++) [23]	3.0B	IN-22K + IN-1K + ext-70M (85M)	O365	✓	64.2
FocalNet-H (DINO) [26]	746M	IN-22K (14M)	O365	×	64.3
Group DETR v2 (Our method)	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 [28] and FocalNet [26]. 'w/ Mask' means using mask annotations when finetuning the detectors on COCO [13].

Abstract

*We present a strong object detector with encoder-decoder pretraining and finetuning. Our method, called Group DETR v2, is built upon a vision transformer encoder ViT-Huge [8], a DETR variant DINO [28], and an efficient DETR training method Group DETR [3]. The training process consists of self-supervised pretraining and finetuning a ViT-Huge encoder on ImageNet-1K, pretraining the detector on Object365, and finally finetuning it on COCO. Group DETR v2 achieves **64.5** mAP on COCO test-dev, and establishes a new SoTA on the COCO leaderboard*[1].

1. Introduction

Recent studies show the effectiveness of transformer models at scale [8, 15, 27]. With encoder pretraining on large-scale data [7,19], the models [1,4,9,18,22,24] are able to achieve superior results on various vision tasks, including object detection. With supervised encoder-decoder pretraining on a large-scale dataset, Object365 [20], DINO [28] achieves a state-of-the-art result on COCO [13].

Our method, *Group DETR v2*, is built upon ViT-Huge, DINO, and Group DETR. We adopt an encoder-decoder pretraining and finetuning pipeline: pretraining and then finetuning a ViT-Huge encoder on ImageNet-1K [7], pretraining the detector, both the encoder and the decoder, on Object365, and finally finetuning it on COCO. Group DETR v2 achieves **64.5** mAP on COCO test-dev [13] (Table 1 and Table 2), setting a new record for COCO object detection. We expect that the results will be further improved with more data and larger models.

2. Method

2.1. Architecture

Encoder. We adopt a ViT-Huge as the encoder. We resort to the self-supervised pretrained model, Vit-Huge, e.g., with the MIM method CAE [4], which shows superior per-

[*] Equal contribution.

[†] Corresponding author.

[1] https://paperswithcode.com/sota/object-detection-on-coco

Table 2. **Our method, Group DETR v2, establishes a new SoTA on the COCO *test-dev* leaderboard.**

Method	mAP	AP50	AP75	AP_s	AP_m	AP_l
Group DETR v2	**64.5**	81.8	71.1	48.4	67.2	77.1

Table 3. **Results on Object365 5k val** with a single scale of 800×1333.

Method	mAP	AP50	AP75	AP_s	AP_m	AP_l
Group DETR v2	**55.6**	68.8	60.9	36.6	57.5	71.3

formance on downstream tasks. We follow ViTDet [12] to build multi-scale feature maps for multi-scale DETR.

Decoder. We adopt the transformer encoder-decoder framework as the decoder that shows promising detection results, including DETR [2], Conditional DETR [17], DAB-DETR [14], Deformable DETR [30], DN-DETR [10], and DINO [28]. Group DETR [3] provides further progress in improving the training convergence speed and the detection performance of various DETR variants. We build our detection decoder by combining DINO [28] into Group DETR [3].

2.2. Implementation

The training process includes three stages: (i) pretrain and finetune the ViT-Huge encoder on ImageNet-1K [7], (ii) pretrain the whole detector (encoder and decoder) on Object365 [20], and (iii) finetune the detector on COCO [13]. When pretraining the detector on Object365, we follow DINO [28] 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 [21], use 100 DN queries [28], and adopt 11 groups with Group DETR [3] during detector pretraining and finetuning. When finetuning the detector on COCO, we find that applying learning rate decay [1,4,5,9] for the components of the detector (encoder and decoder) gives a $\sim$0.9 mAP gain on COCO.

3. Experiments

Results on Object365 5k val. We pretrain Group DETR v2 for 24 epochs with 64 A100 GPUs on Object365. On the first 5k validation set, our Group DETR v2 achieves **55.6** mAP. Table 3 gives detailed results.

Results on the COCO *test-dev*. We finetune the detector (pretrained on Object365) on the COCO training set for 20 epochs with 32 A100 GPUs. During testing, we adopt test time augmentation with various scales and their flipped counterparts, and perform fusion on the query features[2] and the final predictions [28]. Our Group DETR v2 achieves **64.5** mAP on COCO *test-dev*. Table 2 provides detailed results.

Comparisons with state-of-the-art results on the COCO *test-dev* leaderboard. We report the previous SoTA results on the COCO leaderboard. Table 1 shows that only pretraining the ViT-Huge encoder on ImageNet-1K, Group DETR v2 outperforms other methods with larger models (e.g., BEIT-3 [22] and SwinV2-G [15]) and more training data, and sets a new record on COCO *test-dev*. We expect that the results will be further improved with more data and larger models.

References

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

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 2

[3] Qiang Chen, Xiaokang Chen, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, and Jingdong Wang. Group detr: Fast detr training with group-wise one-to-many assignment. *arXiv preprint arXiv:2207.13085*, 2022. 1, 2

[4] 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. *arXiv preprint arXiv:2202.03026*, 2022. 1, 2

[5] 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

[6] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382, 2021. 1

[7] 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. 1, 2

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,

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

Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 1, 2

[10] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13619–13627, 2022. 2

[11] 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. 1

[12] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint arXiv:2203.16527*, 2022. 2

[13] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 1, 2

[14] 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. 2

[15] 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. 1, 2

[16] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 1

[17] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3651–3660, 2021. 2

[18] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022. 1

[19] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 1

[20] 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. 1, 2

[21] 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

[22] 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. 1, 2

[23] 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. 1

[24] Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663, 2022. 1

[25] 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. 1

[26] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. Focal modulation networks, 2022. 1

[27] 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

[28] 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

[29] 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. 1

[30] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2